NOT TO BE FURTHER COPIED OR REPRODUCED

(* y VS. x *)

8

Plots of Relationship

	Exh.	Page
Introduction 8-1	A	
fit, at 8-2	B	
residual, at 8-2	1	8-12
response, at 8-2	2	8-13
factor, at 8-3	3	8-15
	4	8-16
8A. How to plot y against x 8-3	C	
- choice of ruling 8-4	D	
- choice of scale units 8-4	5	8-22
- consequences of our purpose 8-5	6	8-25
- tracing paper 8-6	7	8-26
- kinds of grids 8-6	8	8-28
- shape of plot 8-8	9	8-29
- plotting without graph paper 8-9	10	8-31
	11	8-32
8B. Looking at subtraction 8-10	12	8-33
	E	
8C. Subtracting straight lines 8-18	13	8-37
- finding the line 8-18	14	8-38
- subtracting different straight lines 8-19	15	8-40
	16	8-41
8D. Plotting the population of the U.S.A.	17	8-42
- the later decades 8-24	18	8-45
- coming to details 8-27	H	
- the earlier details 8-30	P	
	19	8-49
8E. Plotting the ratio of births to deaths 8-35	20	8-50
- another try 8-36	21	8-51
- going to the map 8-43	22	8-52
	23	8-54
**********	24	8-55
8H. How far have we come? 8-46		
**********		
8P. Problems 8-48		

8

Plots of Relationship

We have learned something about scratching down batches -- sets of similar numbers. We found more things to do -- saw that more things could be learned about data by doing them -- than we might have expected. It is now time to turn to the use of simple plots, to the plotting of y against x and find, again, more things to do and more gains from doing them than we might expect.

The most quoted passage of all the Sherlock Holmes stories concerns the unusual behavior of the dog in the night time. Holmes called attention to its unusualness -- Watson couldn't see what was unusual -- Holmes pointed out that the dog did nothing.

The moral is clear -- we ought to judge each occurrence against the background of other "nearby" occurrences. We do not ordinarily think of "zero" as unusual -- yet a 24-hour period when no one died anywhere in the world, or a winter with no snow at all in the White Mountains would be recognized by each of us as strikingly unusual.

As in the large, so in the small. Once we have plotted some data, and found out how it behaves generally, our next step is to look at each of its elements -- at each

"point" -- and ask how it seems to deviate from the general behavior of all the "points". To do this it often helps to make a new plot, one that concentrates on such deviations -- in brief, "it often pays to plot residuals".

As we have begun, so will we continue. Here, as well as later, we will be repeatedly engaged in splitting up data according to one version or another of the key relation

observation = fit + residual .

Here the fit is our current description -- always at least somewhat incomplete, always at least somewhat approximate -- of the overall behavior of the data. Each individual observation is split up into a sum of this fit and what is left over, called

a residual .

Residuals are our main tool in going further. They are to the data analyst what powerful magnifying glasses, sensitive chemical tests for blood stains, and delicate listening devices are to a story-book detective. The permeate all sorts of data analysis and appear in many guises.

The most conventional plots follow a simple pattern. There is a quantity which we think of as

a response .

This will be shown on a vertical scale. There is another quantity --

a factor

-- which we think of as possibly explanatory or descriptive. It will be shown on a horizontal scale. The data -- in whole or part -- will consist of pairs of numbers of the form

(factor, response)

which will be plotted as points using these scales.

Most plots -- at least in the popular press -- concentrate on the fit. Either they are to show us that an unsuspected relationship does appear to be there, or to tell us what the relationship is like. Both are good aims, both can be relevant to analyzing data. Plots "in the large" are of use.

For us the most useful plot will be one that might reveal the unexpected or the unobvious. Sometimes a plot "in the large" will do this. Usually, however, it is the plot of residuals that has the greatest use and the greatest impact.

8A. How to plot y against x

Plots are important. Great differences in ease of construction and great differences in effectiveness of use depend on apparently minor, purely technical procedures. This

section's remarks are to get you over most of the elementary hurdles.

* choice of ruling *

If plotting is to be easy, you want graph paper with at least three different thicknesses of ruling:

- light lines for "units"
- medium lines for "fives"
- heavy lines for "tens" .

A wide variety of graph papers can be had with these characteristics. (Some papers have extra heavy lines for "twenties". You will have to learn for yourself whether these help you more than they bother.)

DO NOT use papers ruled in "fours" and "eights", or in "sixes" and "twelves" for plotting data given in ordinary (decimal) numbers. (For monthly data, of course, "twelves" in <u>one</u> direction do help.)

If you want to plot fast, easily, and accurately, avoid quadrille sheets with only one thickness of ruling, and "dime-store" sheets with only two thicknesses. (If you need to save money, see below.)

* choice of scale units *

When you come to plot, you must choose units. Don't try to make one square equal to 3 units, 7 units, 0.03 units,

0.007 units or any such uncomfortable number. Stick to one square being 1, 2, or 5 times a power of ten. (One square = 20,000 or one square = 0.05 are examples that can work out quite well.)

You will find it hard enough to learn to be a fast and accurate plotter with two three-speed gear shifts (1, 2 or 5 across the graph and 1, 2 or 5 up the graph).

If you need to use an abnormal scale -- three units to the square, for example - convert your numbers into plotting units -- as by dividing by 3 -- by slide rule, hand arithmetic, or what have you, BEFORE you start to plot. It will take less time, overall, to make a good graph.

* consequences of our purpose *

Throughout this account we shall be interested in graphs to be looked at, not to be used to find numbers. Our graphs are means for looking at the data, not stores of quantitative information. This means:

- we will want to keep our eyes on the points,
- we will usually NOT connect one point to the "next" one,
- we will want to suppress the rulings of graph paper, at least from our mind's eye,
- we will need only a few numbered ticks along most axes, horizontal or vertical,

- we will want to use symbols large enough to stand out (and if we need more than one kind, different kinds should be both clearly different and -- usually -- almost equally noticeable).

There cannot be too much emphasis on our need to see behavior.

We must play down or eliminate anything that might get in the way of our seeing what appears to go on.

* tracing paper *

One important tool to this end is the use of tracing paper. If we slip a well-printed sheet of graph paper just below the top sheet of a pad of tracing paper, we can plot on that top sheet of tracing paper almost as easily as if it were itself ruled. Then, when we have the points plotted, some boundary or reference lines drawn, and a few scale points ticked, we can take away the graph sheet and look at the points undisturbed by a grid. We often gain noticeably in insight by doing this. (And we have had to pay for a sheet of tracing paper rather than for a sheet of graph paper.)

* kinds of grids *

The use of tracing paper makes it possible for any of us to really arm himself with an easy-graphing capability of broad scope. A tracing pad and one sheet of each of 20

kinds of graph paper puts us in business to do many things. (Actually, it usually seems worthwhile to have a reasonable number of sheets, perhaps in the form of a pad, of at least one or two of the most used kinds of paper. Many will wish to first plot on the graph paper and then copy on the tracing paper.)

"Ordinary" graph paper frequently comes 8 to the inch (beware, it may be in "4's" or "8's"), 10 to the inch, 12 to the inch (beware, it may be in "6's" or "12's"), 20 to the inch and 10 to the centimeter. The writer likes the coarser ratings at least as well as the fine; it is up to you to learn your own preference. (Beware of 12 to the inch rulings, since too many are ruled in "sixes" and "twelves"; but bear in mind that 12 to the inch ruled in fives and tens -- like Codex Nos. 31,253 and 32,253, for example, is a very good base for graphs where the final product is a typed page. It has little special advantage for most other purposes.) Never use 8 to the inch ruled in 4's and 8's (unless what you are plotting comes in 1/8th of an inch, or, like stock market prices, in 1/8ths of a dollar).

Semilogarithmic (uniform scale one way, log scale the other) and full logarithmic (log scales both ways) graph papers come in a variety of scales and patterns. To have a reasonable selection is to save time and encourage inquiry.

One CAN always look up the logs and then plot on uniform-scale graph paper, but WILL one? If two-decimal logs will do, and chapter 3 is at hand, probably yes. Otherwise? As we saw above, logs are often the way to make data reveal itself.

A variety of other graph papers are useful, as shall see later. We note here that paper with two square-root scales is available (Codex No. 31,298 or No. 32,298) and that trilinear or isometric paper -- with three sets of rulings at 120° to one another -- is often quite useful. (It is available from various manufacturers.) Reciprocal (one way) paper is available from both Codex and K & E.

* shape of plot *

Naturally enough, graph paper comes in the same general shape as ruled paper (and writing paper). Clearly it can be used in two ways:

- narrow edge at the top of the plot, or
- broad edge at the top.

By analogy, with the use of ruled paper it is natural to make plots with the narrow edge up. Full use of the area then usually makes a plot that is taller than it is wide.

There are some purposes for which this is a good shape. The ever-faster rising curve of early growth is an example. We can quite well use taller-than-wide plots for such simple pictures, most of which tell us that we haven't

yet made an analytically useful plot. Sometimes, indeed, such simple plots can be clearer when they are taller than wide rather than wider than tall.

Most diagnostic plots either involve a more or less definite dependence that bobbles around a lot or a point spatter. Such plots are rather more often better made wider than tall. Wider-than-tall shapes usually make it easier for the eye to follow from left to right. For still other purposes, taller-than-wide plots are to be preferred.

When a plot is made wider than tall, convention says it should be turned in the direction illustrated by many plots in this chapter, even if this makes a down-to-up legend upside down when we first see the plot as we turn the pages.

* plotting without graph paper *

We almost always want to look at numbers. We do not always have graph paper at hand. There is no excuse for failing to plot and look.

We usually have ruled paper at hand. For emergency graph paper, take out one sheet of ruled paper, turn it on its side, and place it beneath another sheet of ruled paper. If these two sheets have a light-colored backing -- often provided by the rest of the pad or notebook -- the vertical lines on the lower sheet are almost certain to show through well enough,

combining with the horizontal lines on the top sheet to form a grid on which plotting is reasonably easy. (The first step in this sort of plotting is to mark -- by ticks or unobtrusive dots -- enough information on the top sheet to make it easy to get the lower sheet back to its original position after it slips.)

With this technique, one can make useful, if not decorative, plots almost anywhere.

8B. Looking at subtraction

Undoubtedly, the form of graphical representation we have been most exposed to in school is one involving two variables called x and y, in which y is said to depend on x. In data analysis, a plot of y against x may help us when we know nothing about the logical connection from x to y -- even when we do not know whether or not there is one -- even when we know that such a connection is impossible.

Before we can make full use of such plots, we need to understand -- both in terms of doing and of feeling -- certain things about such a plot, including:

- how to subtract one "curve" from another.
- how to find numerical expression for a straight line drawn on graph paper.
- what effect subtracting different -- two or more -- straight lines from the same data points has.

- how to try to re-express either y or x or both so as to make the data appear more nearly straight.

- why graph paper expresses the essence of the situation much better than the kind of picture we will usually want to look at.

When we have x alone, subtraction is simply and easily represented by sliding arrows along. Subtracting 3 from 5 is a matter of drawing an arrow starting at +3 and extending to +5, and then sliding this arrow until it starts at 0. Its new end point, +2, is the result of subtraction. Exhibit 1 shows this example and three others that involve various combinations of minus signs.

In dealing with data we usually deal with subtraction of y's rather than of x's. For a hypothetical ABC Corporation, in 1960, sales were 44 million and expenses were 32 million. We can easily subtract expenses from sales to find profit before taxes of 12 million. Again we are sliding the arrow to put its base at zero. This time we have to be careful to keep it at 1960.

In this exhibit and the next -- for clarity of what we are doing, rather than for clarity of result -- we have used two time scales placed side by side. Ordinarily we would use one time scale and slide each arrow along a single vertical line -- as for the broken arrows on the left sides of these exhibits.

exhibit 1 of chapter 8

Four Examples of Subtraction

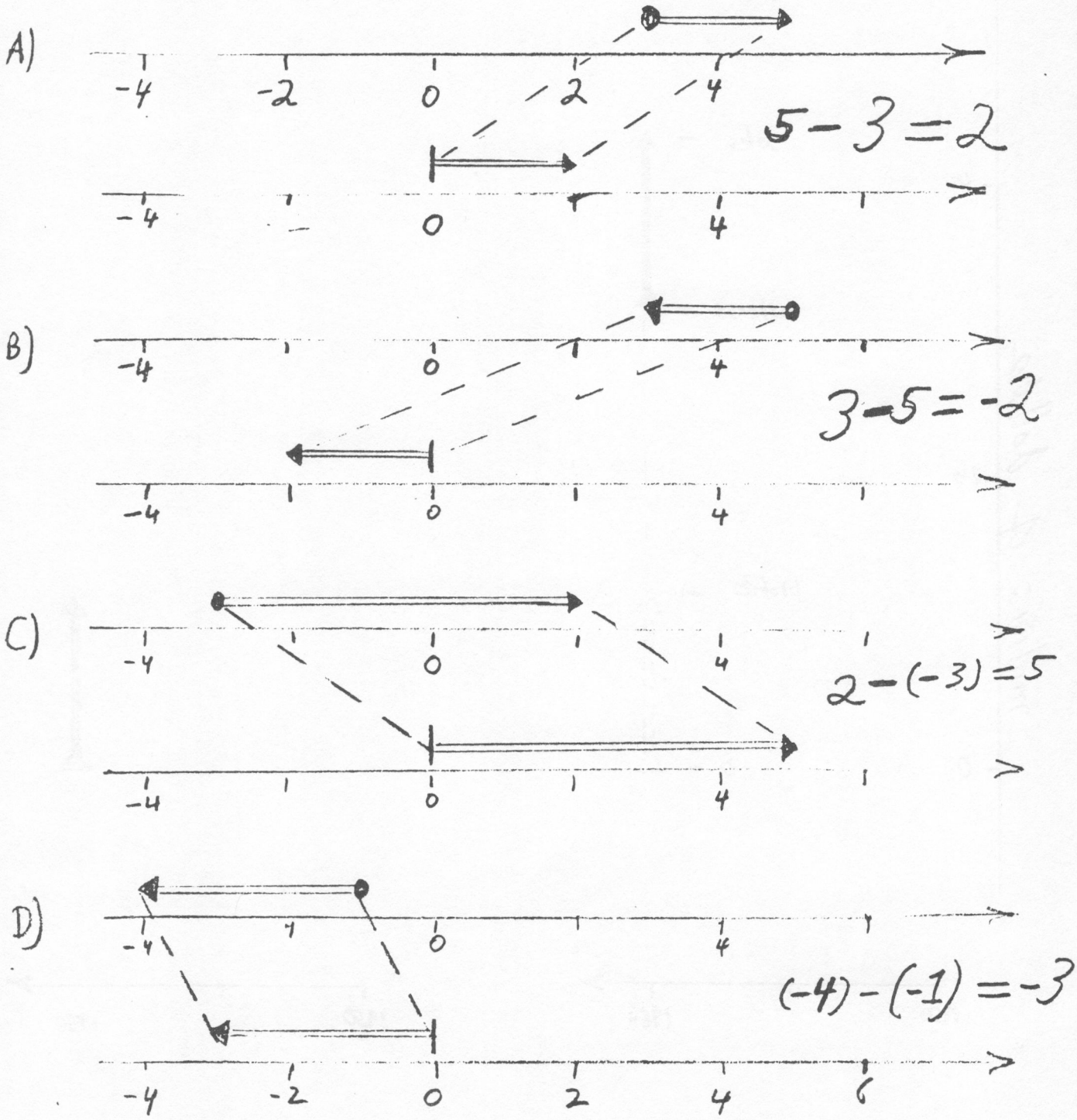


exhibit 2 of chapter 8

The ABC Corporation in 1960

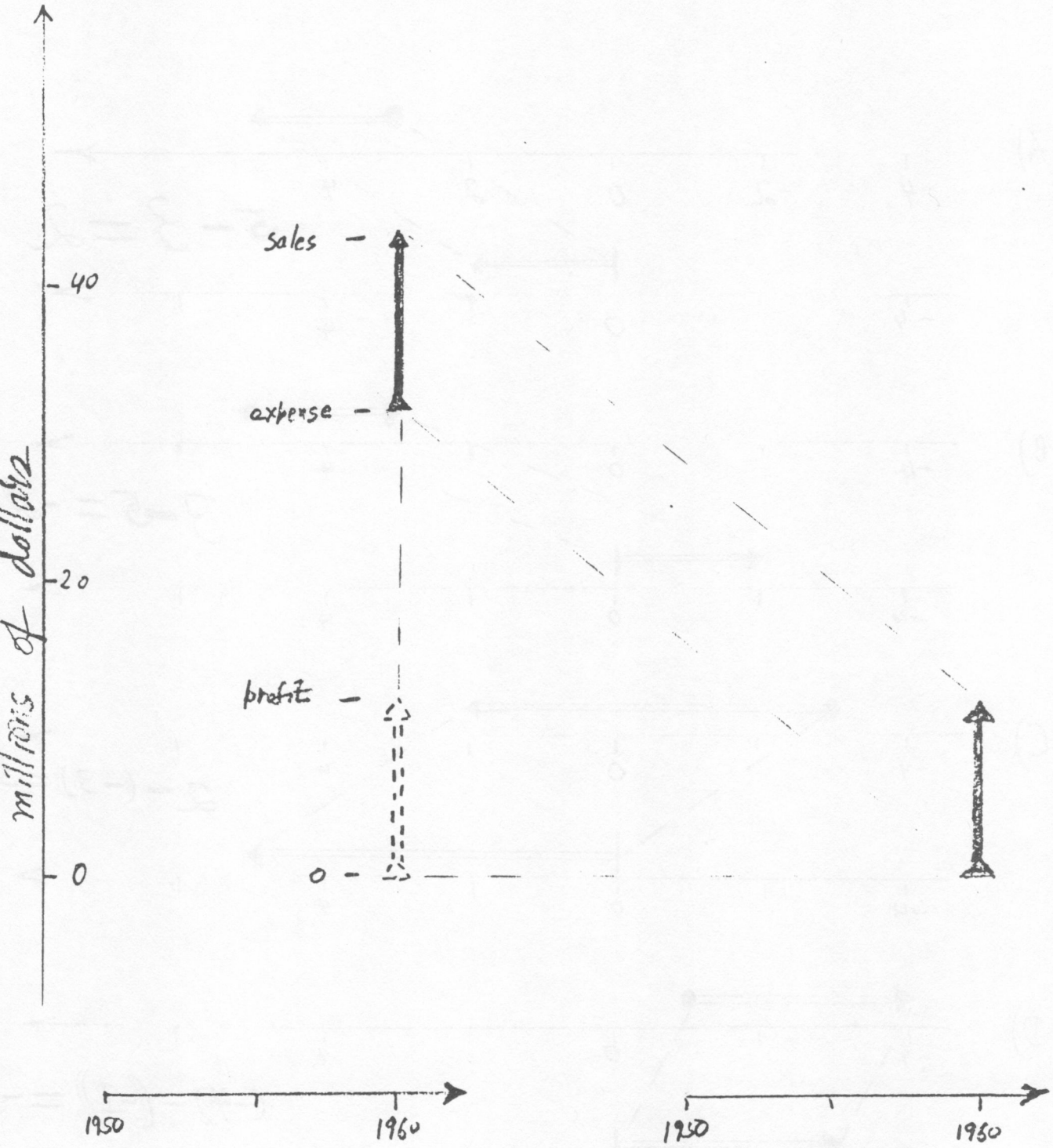


Exhibit 3 retains the pattern and shows sales, expenses, and profits for 12 consecutive years. To avoid confusing detail, the arrows and their sliding are shown for only three years, 1951, 1957, and 1960.

The case where we will be most concerned with graphical subtraction is the case where we solve the basic relation

data = incomplete description + residual
= fit + residual

for the residual, finding

residual = data - incomplete description

Exhibit 4 shows an example of this where the incomplete description is a straight line. Here we have slid the vertical arrows down (or up) the vertical line on which they lie. Again we have shown only three arrows of the many possible.

It is natural to think -- and speak -- of such a subtraction of a straight-line partial description as an untilting. Natural and useful. Yet such use of words seems likely to distort our thinking a little if we are not wary. Untilting suggests, at least to some of us, a rigid motion in which the final curve comes from the initial curve by a rigid motion. This is quite wrong.

exhibit 3 of chapter 8

Twelve Years of the ABC Corporation

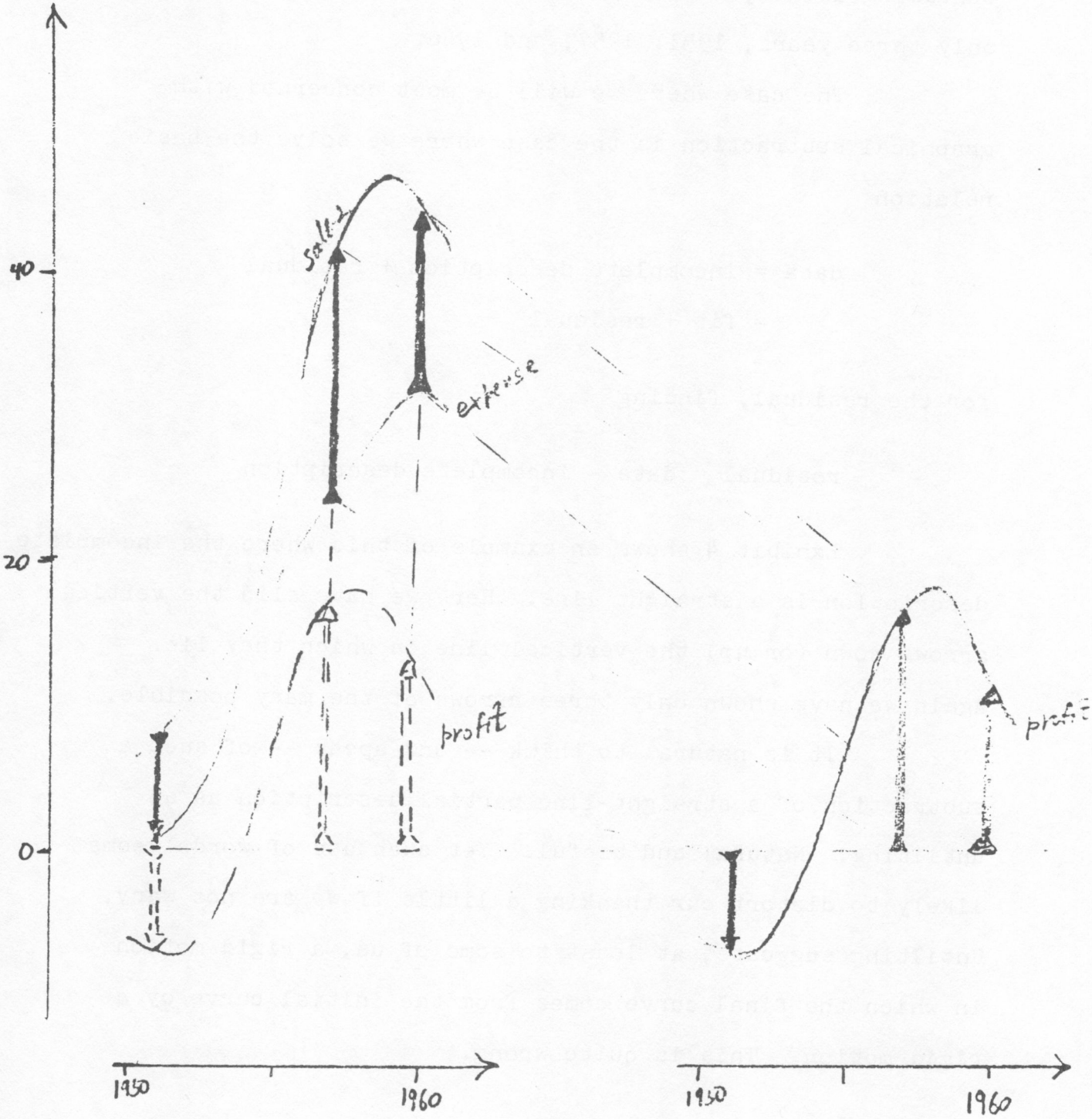
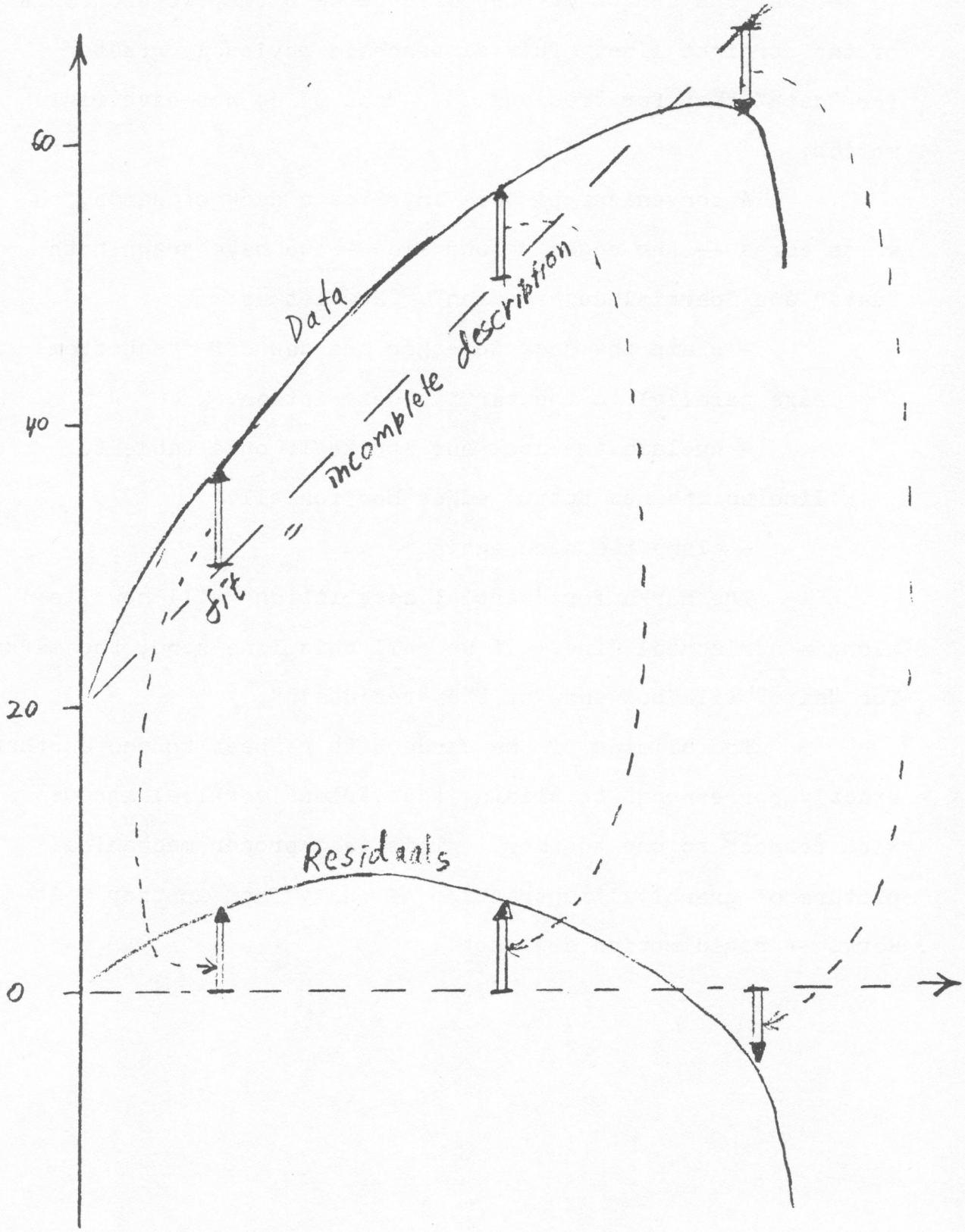


exhibit 4 of chapter 8

Subtraction of An Incomplete Description From Data to Form Residuals



There are many ways to see this. An easy one is to look at the length of the curve between the two crossings of the straight line. This distance is obviously greater for "data" than for "residuals". Thus we do not have rigid motion.

A convenient picture involves a deck of cards, on whose edges -- the edges on one side -- we have drawn both "data" and "partial description". Now let us:

- clamp the deck together and saw off its bottom edge parallel to the partial description.
- unclamp the deck and strike it on a table to line up its new bottom edges horizontally.
- clamp the deck again.

The marks for "partial description" will now lie along a horizontal line. If we call this line zero, the marks for "data" will now show us the "residuals".

The sliding of the cards with respect to one another exactly corresponds to sliding the various vertical arrows with respect to one another. This is a proper mechanical picture of graphical subtraction of one y from another. It works -- rigid motion does not.

8C. Subtracting straight lines

If we have taken our data, plotted it, and drawn a straight line through it, and now we wish to use the straight line as an incomplete description, our next task is to subtract the line from the data. Sometimes, as in exhibit 4, we can do the subtraction graphically. Often, however, this is more work than we like.

* finding the line *

To do the subtraction by arithmetic, we have to turn the straight line into numbers. The easy way to do this is to choose two points on the line -- let us call them (x_1,y_1) and (x_2,y_2) -- read off their coordinates, and say that an equation for the line is

$$y = y_1 + b(x-x_1)$$

where the slope, b, of the line is given by

$$b = \frac{y_2-y_1}{x_2-x_1} \quad .$$

Clearly, where $x = x_1$, we have $x - x_1 = 0$ and $y = y_1$. When $x = x_2$, the second term on the right-hand side of the equation is

$$\frac{y_2-y_1}{x_2-x_1}(x_2-x_1) = y_2-y_1$$

and we have

$$y = y_1 + (y_2 - y_1) = y_2$$

as we should.

To make the arithmetic of using the equation easy, we should choose x_1 so that the values of $x - x_1$ are as simple as possible. To make finding the equation easy, we should choose x_2 so that $x_1 - x_2$ is a simple number. There need never be great difficulty in turning a straight line into an equation.

* subtracting different straight lines *

If we look at data, and draw what looks like a reasonable straight line through it, we are not likely to draw exactly the line that will serve us best. What are the consequences of a somewhat unsatisfactory line?

- what will it do to the residuals?
- how hard will it be to move to a better choice?

An easy way to approach both these questions is to ask what happens if we subtract first one straight line from y and then a second straight line from the residuals thus formed. Algebra is easier than geometry here. If we begin by subtracting a + bx we will form

$$y - (a+bx)$$

If from this we subtract A + Bx, we will form

$$[y - (a+bx)] - (A+Bx) = y - [(a+A) + (b+B)x] .$$

The result of two subtractions -- of a + bx and A + Bx -- is always the same as the result of a single subtraction

$$(a+A) + (b+B)x$$

-- of a subtraction of the

sum of the two lines .

Those who wish can show the same fact geometrically.

This fact about subtracting lines goes far toward answering our two questions:

- if we subtract one line, and on looking at the residuals find them still tilted, we are free to draw a line among these residuals and subtract it further, thus finding new residuals. The new residuals correspond to subtracting a single line -- the sum of the two actually subtracted. We need not go back and start again. This is particularly handy when the residuals are much smaller numbers than are the data.

- if we subtract an unsatisfactory line, and discover that we have done this by looking at the residuals, we could always correct this by a further

subtraction. Accordingly, our first residuals will differ from the better residuals by being somewhat tilted. Since slight tilts do little to hide what we are looking at the residuals for -- evidence of further structure or of unusual values -- it will rarely be necessary to do the second subtraction (unless we want to publish the residuals). We can see what we need to see in the slightly tilted residuals. And, if we want to find the equation of the better line, we can draw a correction line and sum the expressions of the original and correction lines.

Subtracting lines is simple and convenient in many ways.

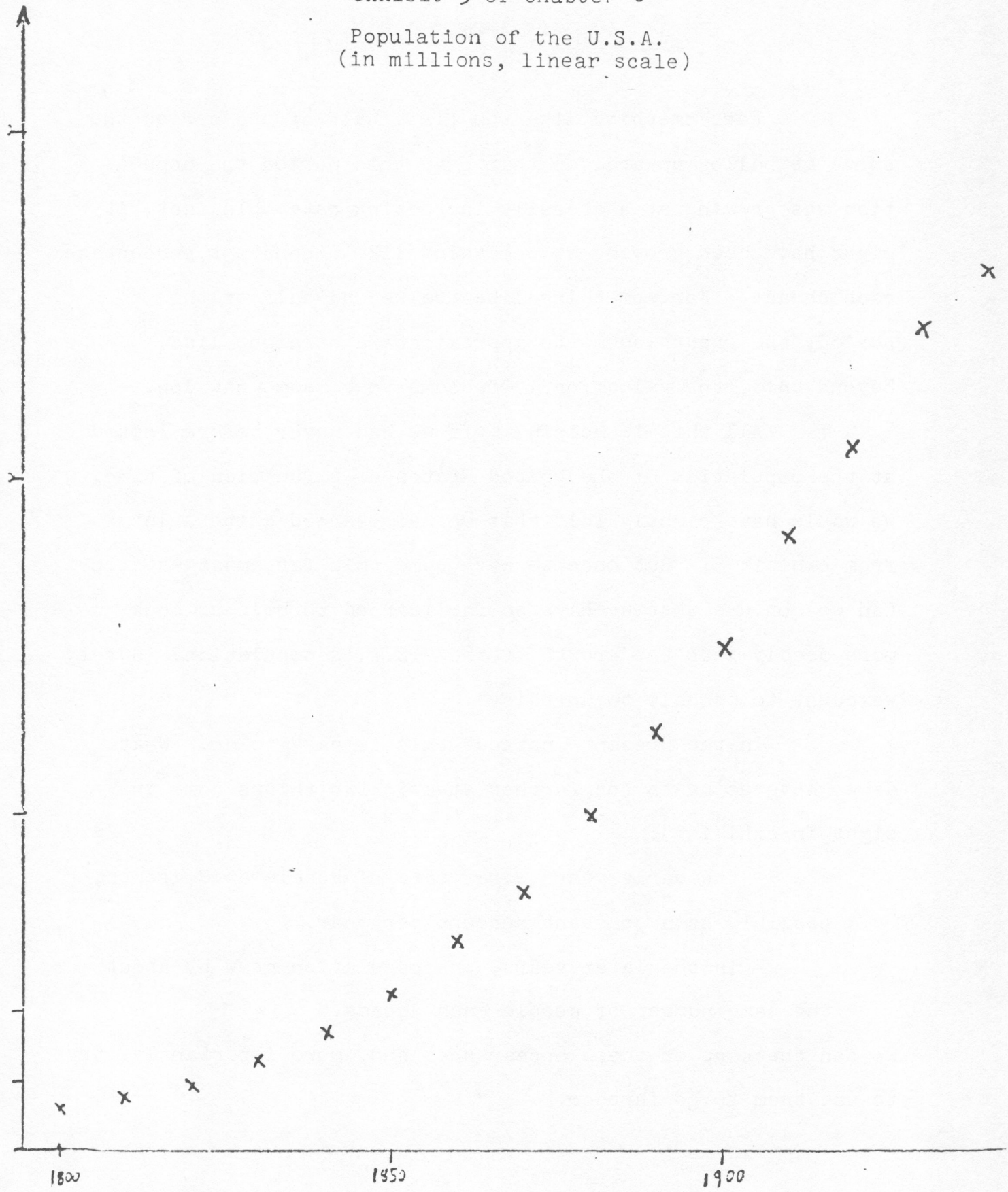
8D. Plotting the population of the U.S.A.

Many people would think that plotting y against x for simple data is something requiring little thought. If one only wishes to learn little, it is true that a very little thought is enough, but if we want to learn more, we must think more.

A convenient example of this is given by the population of the U.S.A. as counted by the census every ten years from 1800 to 1950. Exhibit 5 shows the "little thought" version, in which millions of people are plotted against date. What can we see in this graph?

exhibit 5 of chapter 8

Population of the U.S.A.
(in millions, linear scale)



For something like the first half of the period the curve is hollow upward, so that over this period the population was growing at a steadily increasing rate. In fact, it might have been growing at something like a constant percentage each decade. For something like the second half of the period, the growth seems to approximate a straight line. Beyond this, the value for 1940 seems to be somewhat low.

All this is helpful. If we had never before looked at the population of the United States as a function of time, we would have rightly felt that we had learned quite a lot from exhibit 5. But once we have come this far, must we stop? Can we not use what we have so far learned to help us look more deeply into the growth of the U.S.A.'s population? Surely we ought to be able to do this.

In the present instance this is easy to do. What do we have as bases for further steps? Two things come in sight in exhibit 5:

- the early years were years of accelerated growth, possibly at a constant percent per year.
- in the later years, the population grew by about the same number of people each decade.

We can check up on these appearances and, more importantly, try to use them to go further.

To check up on constant percent growth, the easy thing is to look up the logs of the population sizes, and then plot them against date. (If we didn't expect to go further, we could just plot the raw values on semilogarithmic graph paper instead.) Exhibit 6 shows the result of doing this. The earlier part of the plot now looks quite straight, even if we put our eye close to the paper so that we can look right along this hypothetical line. This looks very much as if constant percentage growth per decade was a good description of U.S.A. population growth in the early 1800's. Let us record this appearance, and plancto come back to it.

* the later decades *

Before going further with the early decades, however, we shall turn to the apparent linearity of population growth seen in the original graph, exhibit 5, for the later decades. Exhibit 7 shows the result of drawing in a comparison line. To find an equation for this line, note that in 1870 the line has a height of about 35 -- 35 million people -- and in 1950 it has a height of somewhat less than 150, say 147. The line through (1870,35) and (1950,147) has slope

$$\frac{147 - 35}{1950 - 1870} = \frac{112}{80} = 1.4$$

and thus equation

$$y = 35 + 1.4(x-1870) .$$

exhibit 6 of chapter 8

Population of the U.S.A.
(in millions, logarithmic scale)

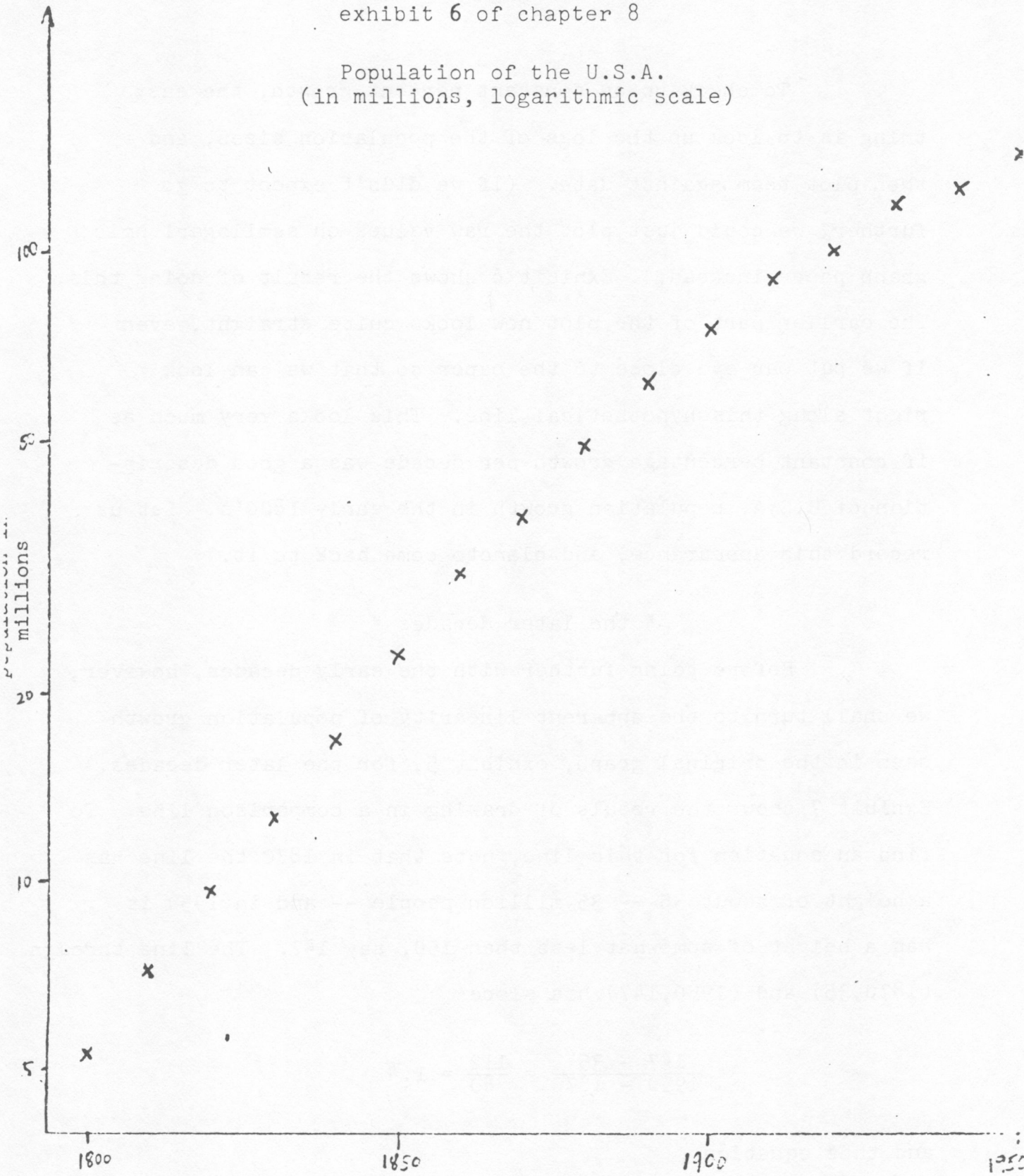
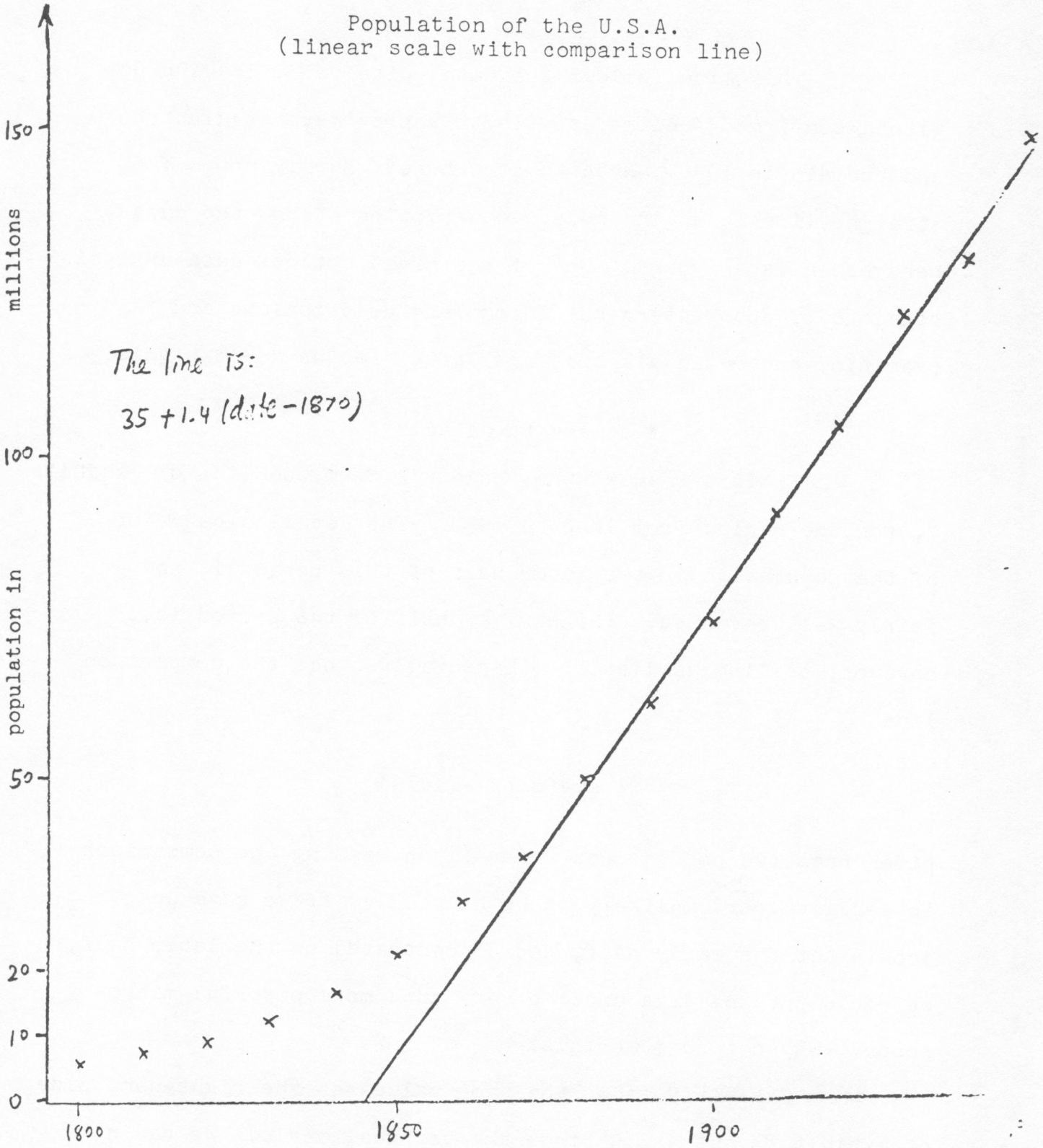


exhibit 7 of chapter 8

Population of the U.S.A.
(linear scale with comparison line)



This line follows the data quite well, leaving us with a confirmed feeling about both approximate straightness and the dip in 1940. Need we stop here? Surely not. The straight line is an incomplete description of how the data behaves in later years. One of the great arts of data analysis consists of subtracting out incomplete descriptions and examining the residuals that are left. Let us do just this.

* coming to details *

Exhibit 8 shows the residuals from the line of exhibit 7, for the period from 1800 to 1950. The detailed behavior of the population in the later half of this period is now fairly well revealed. The earlier half of the period is, however, telling us little. (Especially since the comparison line

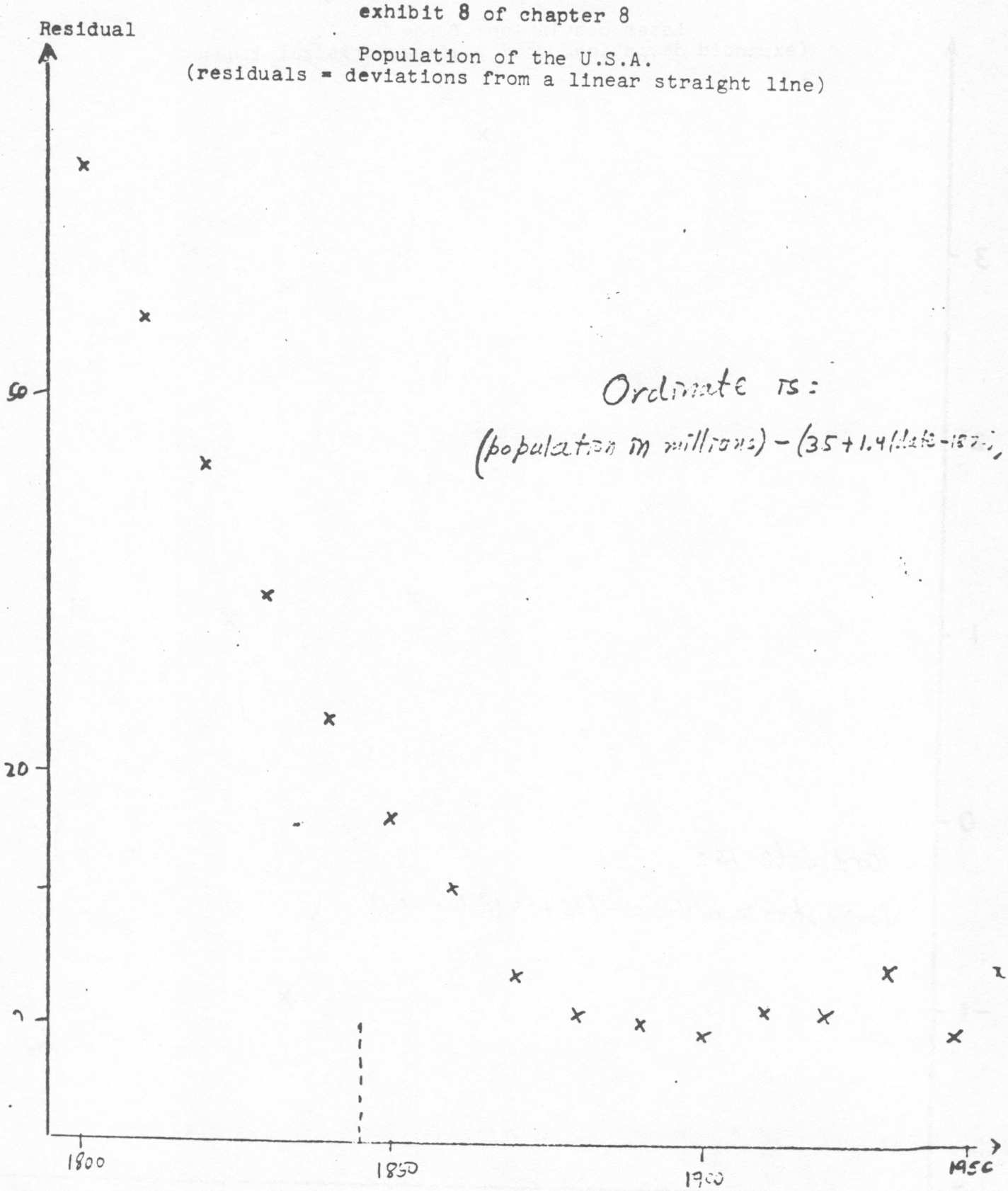
$$35 + 1.4(\text{date} - 1870)$$

gives negative populations before 1845, making the comparison in earlier years unlikely to be helpful.) If we gave up looking at the early half, and concentrated on the later half, we could put the data under a very much more powerful microscope. Why not do just this?

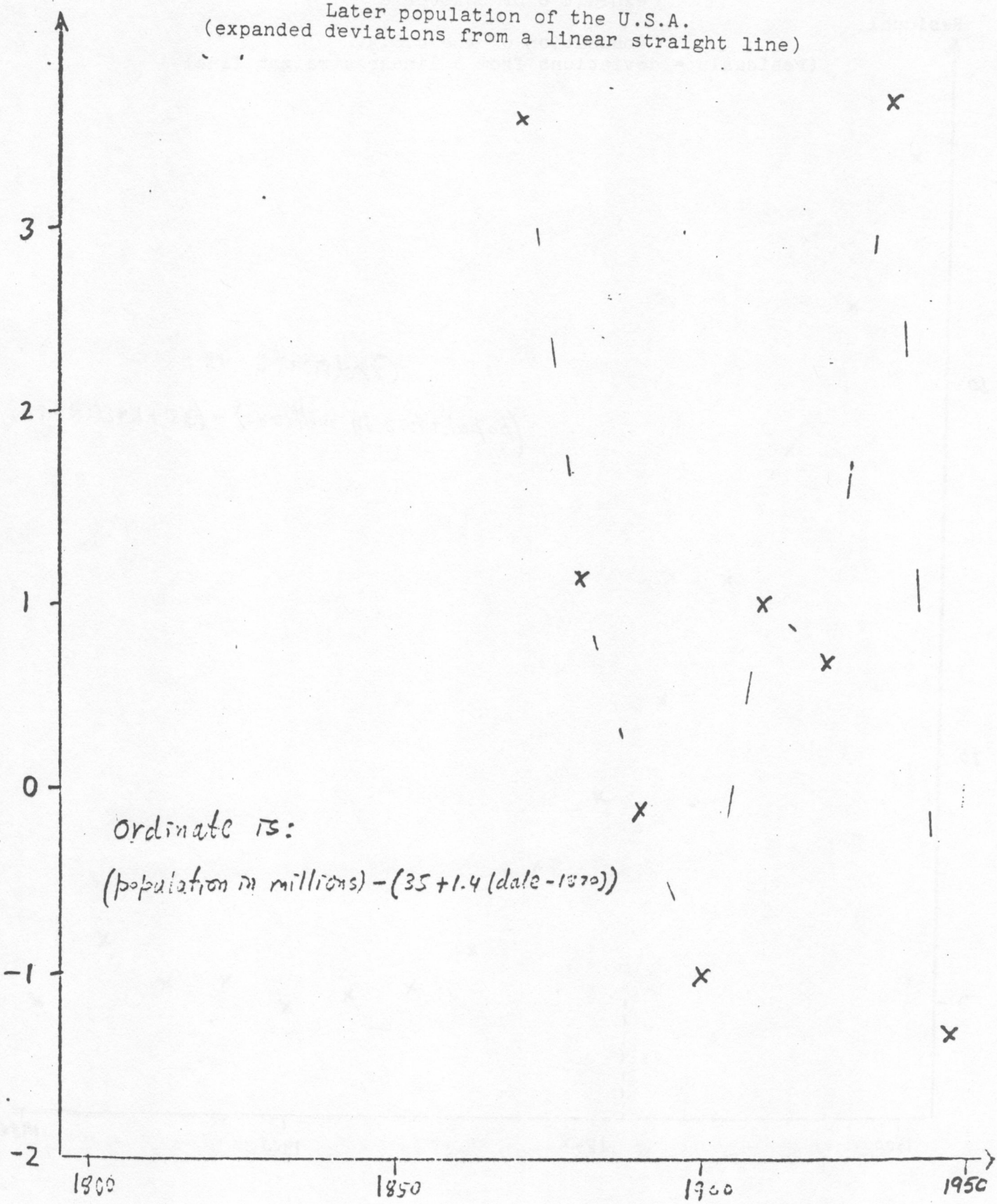
Exhibit 9 shows the same values as the right-hand side of exhibit 8, but at 15 times the vertical scale. We can now see that the 1940 population was about five million less than

exhibit 8 of chapter 8

Population of the U.S.A.
(residuals = deviations from a linear straight line)



Later population of the U.S.A.
(expanded deviations from a linear straight line)



would fit smoothly into the adjacent values. (Why do you think this happened?) And more, for we can now see that 1920 was a couple of millions low also (unless 1930 is thought to be unduly high).

Our magnifying glass is now working at full capacity, at least until we identify some further partial description and arrange to subtract it out also. To learn more about the later years of U.S.A. population growth, we would need either to get year-by-year estimates or to study the mechanisms which are involved.

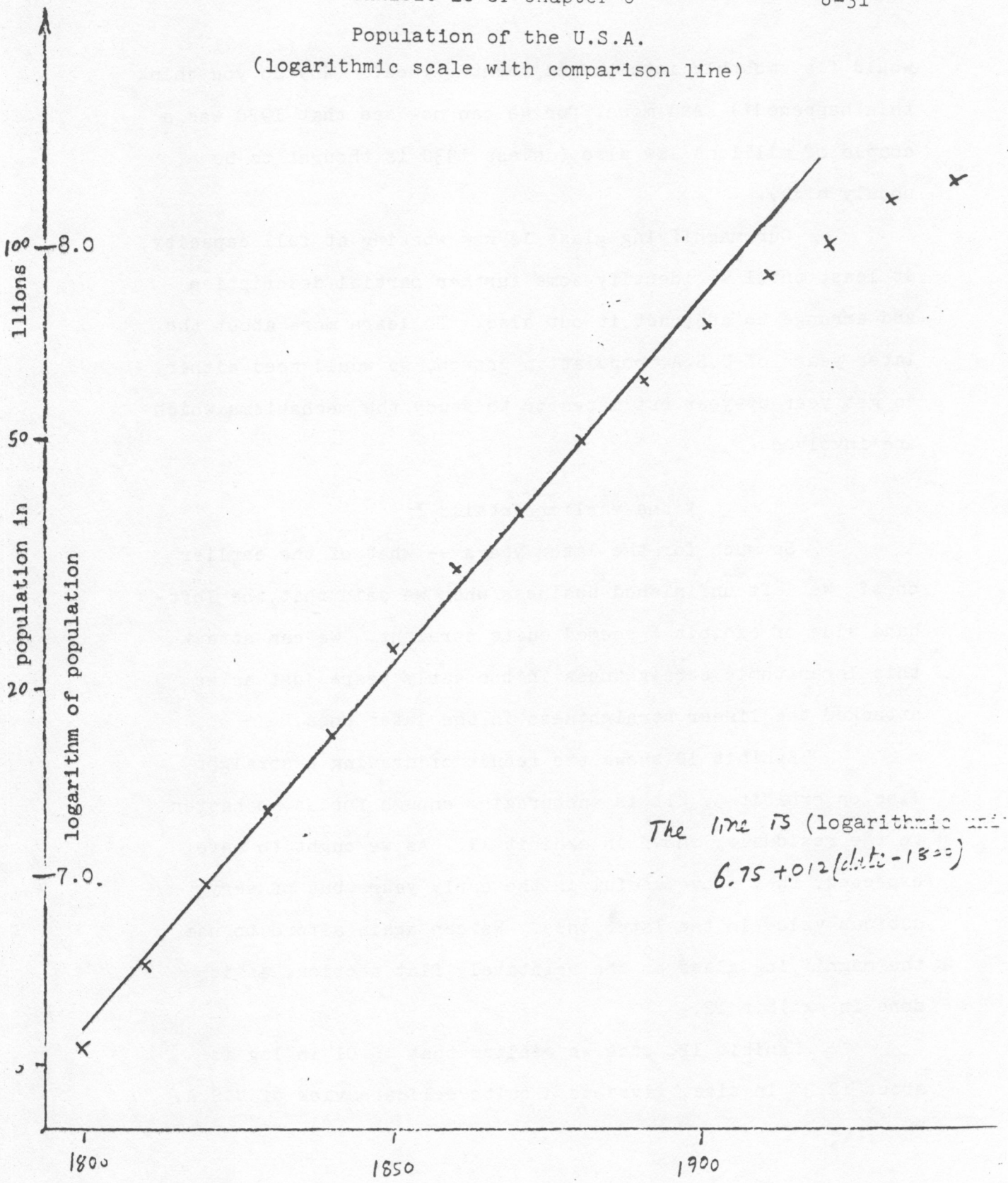
* the earlier details *

So much for the later years -- what of the earlier ones? We left unfinished business when we said that the left-hand side of exhibit 6 seemed quite straight. We can attack this logarithmic straightness in the early years just as we attacked the linear straightness in the later ones.

Exhibit 10 shows the result of drawing a straight line on exhibit 5. It is encouraging enough for us to hasten to the residuals, shown in exhibit 11. As we ought to have expected, they prove useful in the early years but of very dubious value in the later ones. We can again afford to use the magnifying glass on the relatively flat section, as is done in exhibit 12.

Exhibit 12, once we realize that ±0.01 in log is about ±2.3% in size, gives us a quite delicate view of U.S.A.

Population of the U.S.A.
(logarithmic scale with comparison line)



Population of the U.S.A.
(deviations from logarithmic straight line)

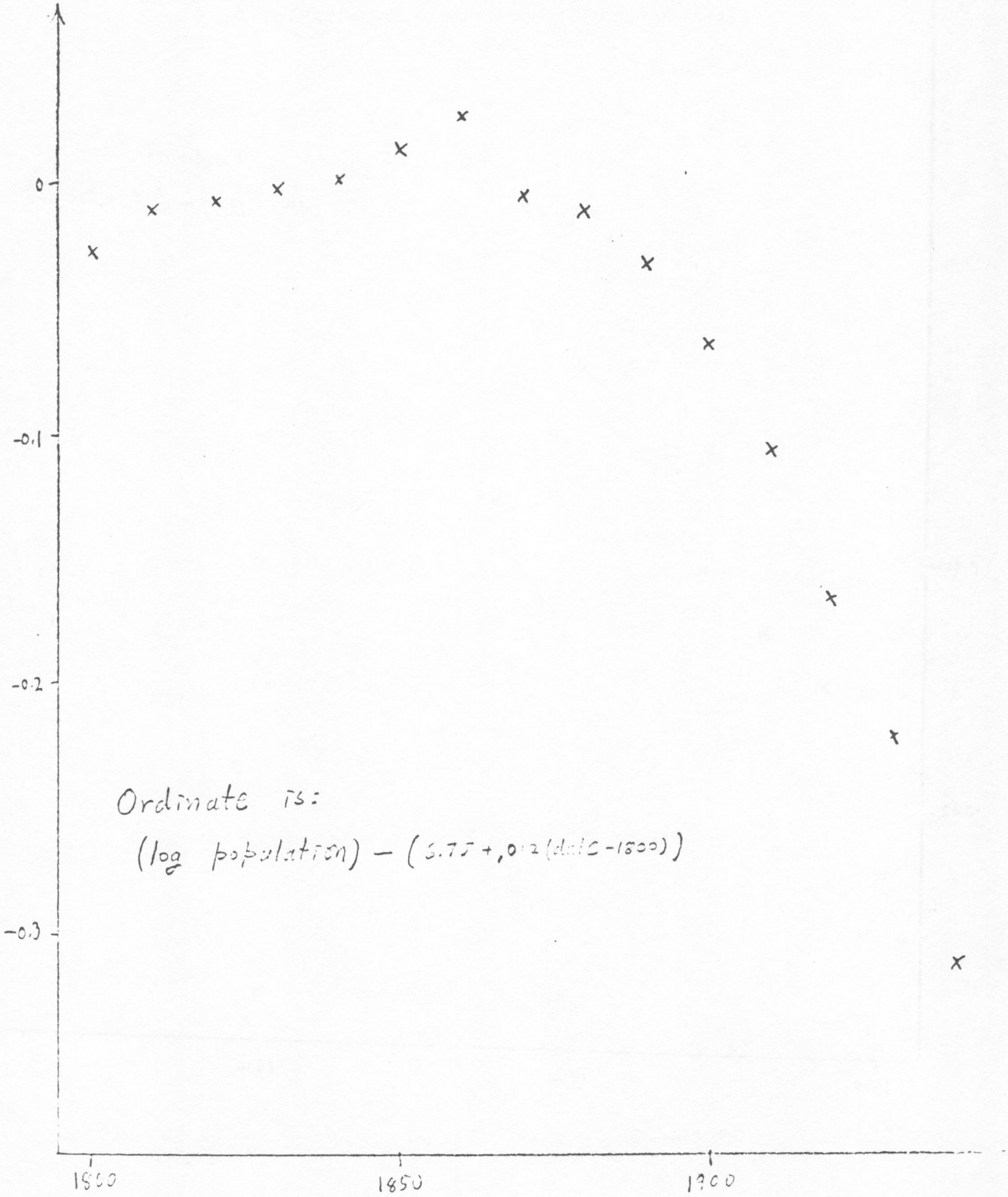
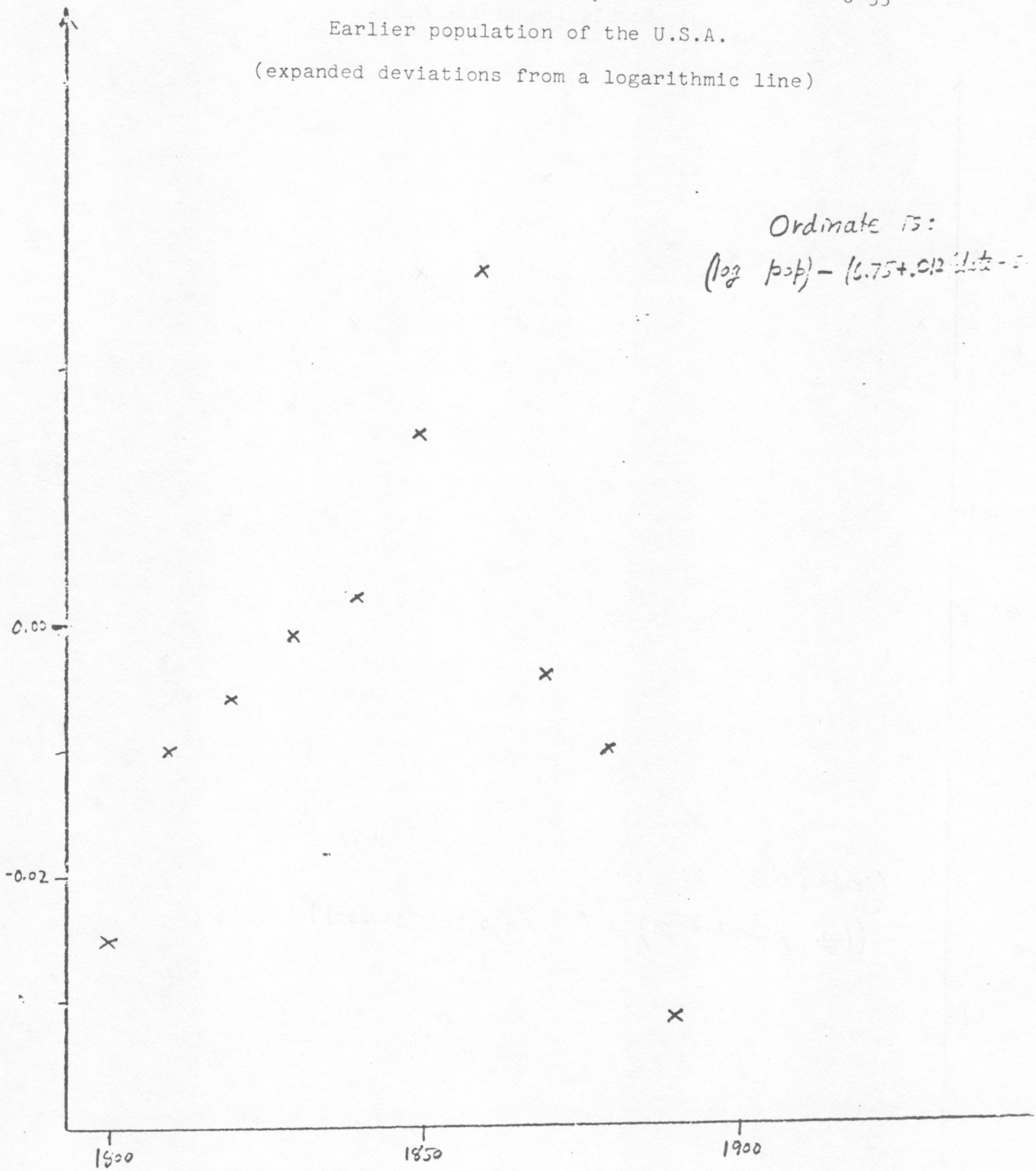


exhibit 12 of chapter 8

Earlier population of the U.S.A.

(expanded deviations from a logarithmic line)



population growth during the 19th century. After 1860, population growth was not as fast as before. Moreover, 1800 appears to have been additionally depressed by 3 or 4%. Why? Again we have gone as far with our microscope as seems reasonable without further inputs.

Experts believe many of the fluctuations now so clear to us are due to variations in completeness of the census, rather than to changes in population growth. Clearly the data cannot contradict the experts. The fluctuations are in the numbers, whatever their source. It is worthwhile to find them, whether they tell us about the growth of U.S. population or the deficiencies of U.S. censuses.

If we had to choose a set of graphs to summarize U.S.A. population growth as completely as we now can, we would probably choose to show all four of the following:

- exhibits 10 and 7, to show the general patterns of growth.
- exhibits 12 and 9, to show local behavior.

Together, these four would be responsive to the request: Make useful plots of U.S. population against date. (If we were population specialists, we would know about logistic functions and be able to fit a single incomplete description all the way from 1800 to 1950. This would simplify the plotting of a single graph of residuals, and would probably allow us to summarize the situation in two plots, one showing the fit and the other the residuals.)

What are the lessons to be learned from this example? Not merely that thought can help us see deeper. We have seen specific example of very general principles, including these:

- choosing scales to make behavior roughly linear always allows us to see local or idiosyncratic behavior much more clearly.
- subtracting incomplete descriptions to make behavior roughly flat always allows us to expand the vertical scale and look harder at almost any kind of behavior.

Whatever the data, we can try to gain by straightening or by flattening. When we succeed in doing one or both, we almost always see more clearly what is going on.

8E. Plotting the ratio of births to deaths

The Census's County and City Data Book contains much varied information. In particular, the 1962 edition gives for each state the number of live births for 1959, the number of deaths for 1959, and the density of population for 1960. One who believed in the "wide-open spaces" might feel that the ratio of births to deaths would appear to be influenced by the density of population, at least if the South and the Atlantic coast states were set aside. A plot of the ratio of births to deaths against population density for the remaining states

appears as in exhibit 13. About all we can say from this plot is that the point spatter seems crudely L-shaped. The use of a linear scale for population per square mile has squeezed so many of the states up against the vertical axis that we can't be sure what is going on. If we are to see what, if anything, is going on, we must adjust the left-to-right scale so that the states are less jammed together.

Exhibit 14 shows the result of using a log scale for population density. We now see the three states with unusually high ratios of births to deaths have low, but not very low densities, and do not appear to be typical of very low-density states. Looking only at the other states there may be a faint tendency for a higher ratio of births to deaths to go with lower population density. (By setting aside the three states, we could double the vertical scale for the rest, putting this tendency under more of a microscope. Doing this teaches us almost nothing new, as the reader may verify.)

* another try *

If we are to give up on population density, what next? The three unusual states (and their birth/death ratios) were: New Mexico (4.95), Utah (4.46), and Arizona (3.83). Looking again in the <u>County and City Data Book</u>, we discover that all three have young populations, for instance, as measured by median age. Thus it is natural to plot births/deaths against median age.

exhibit 13 of chapter 8

Births/deaths and population density

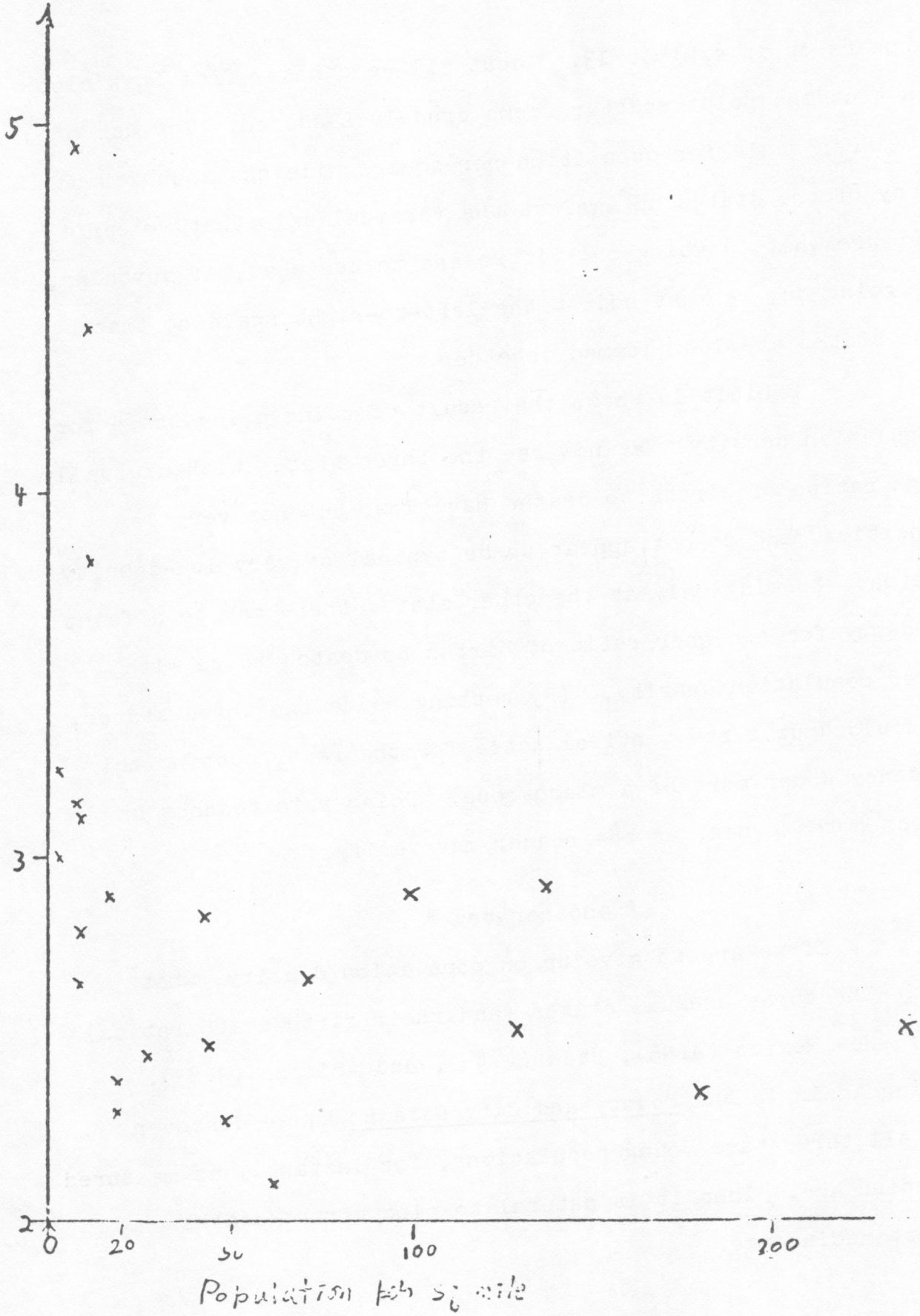


exhibit 14 of chapter 8

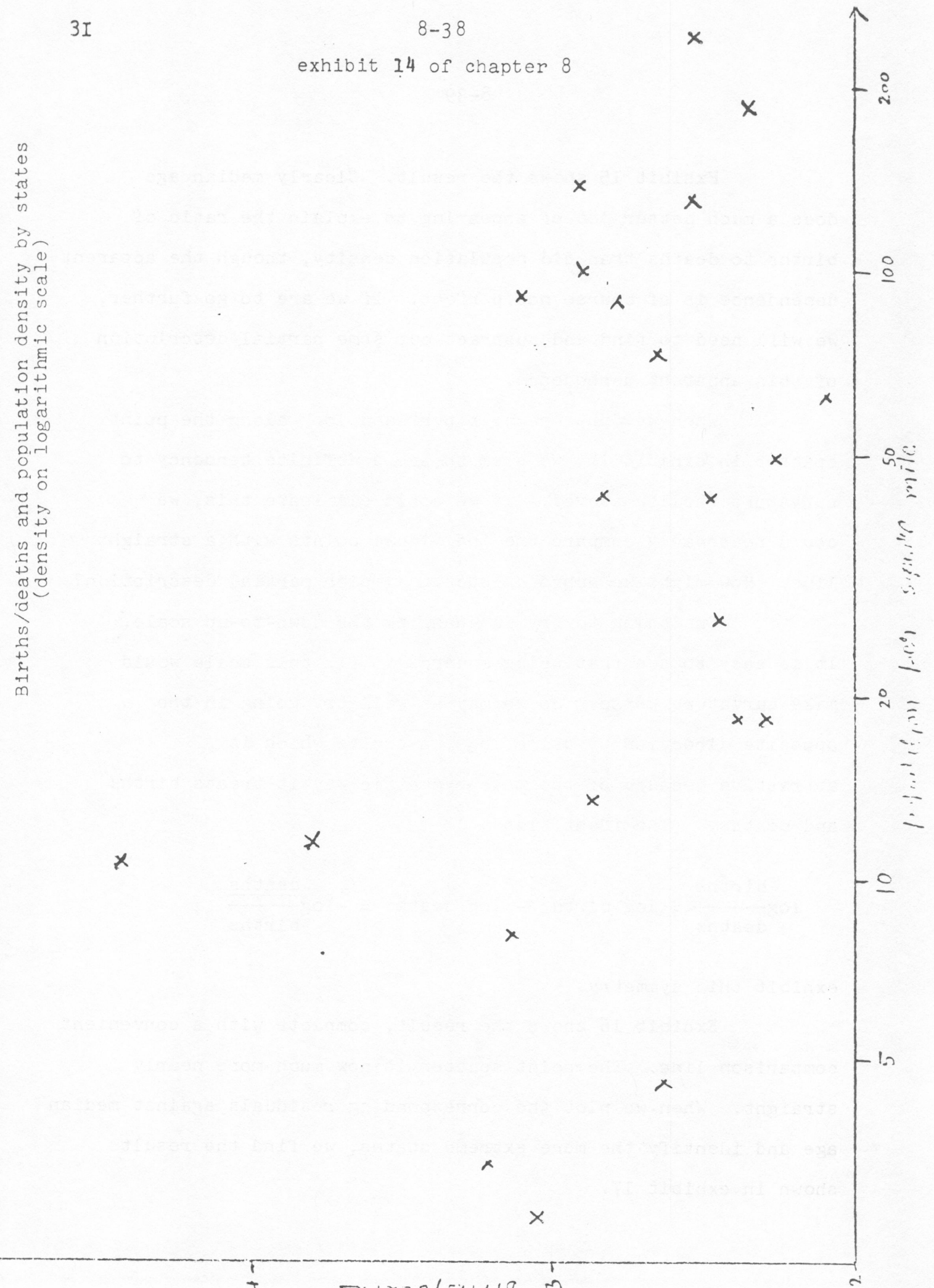


Exhibit 15 shows the result. Clearly median age does a much better job of appearing to explain the ratio of births to deaths than did population density, though the apparent dependence is of course not perfect. If we are to go further, we will need to find and subtract out some partial description of this apparent dependence.

When we "put eye to paper" and look along the point spatter in exhibit 15, we seem to see a definite tendency to curvature (hollow above). If we could eliminate this, we could reasonably compare the individual points with a straight line. How might we approach such a simpler partial description?

One thing to try is changing the down-to-up scale. It is easy to see that using squares to fix this scale would make curvature worse. So we may as well try going in the opposite direction by using logs, a choice which is attractive because of the more symmetric way it treats births and deaths. (The identities

$$\log\frac{\text{births}}{\text{deaths}} \equiv \log \text{births} - \log \text{deaths} \equiv -\log\frac{\text{deaths}}{\text{births}}$$

exhibit this symmetry.

Exhibit 16 shows the result, complete with a convenient comparison line. The point spatter is now much more nearly straight. When we plot the corresponding residuals against median age and identify the more extreme states, we find the results shown in exhibit 17.

exhibit 15 of chapter 8

Births/deaths and median age by state

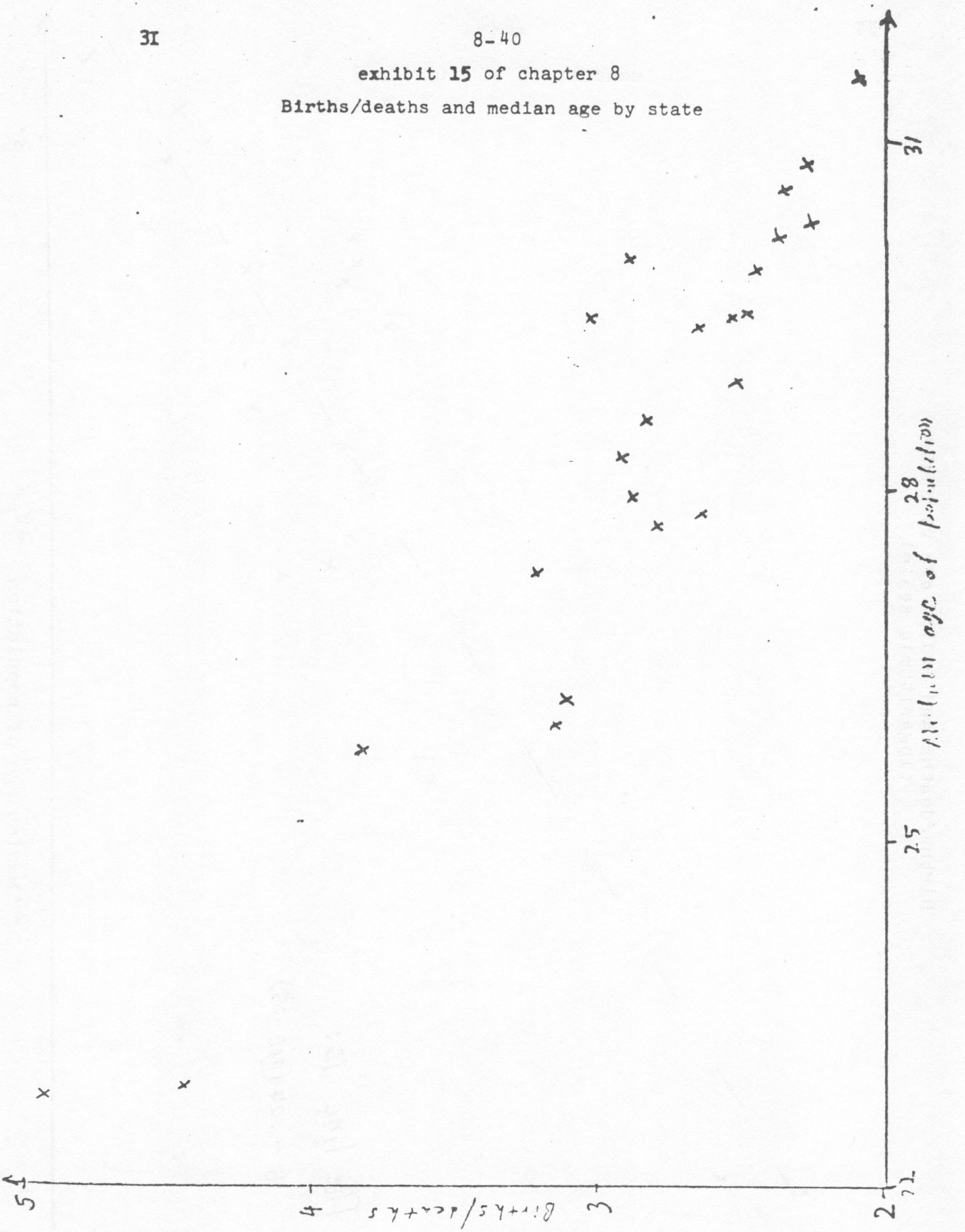
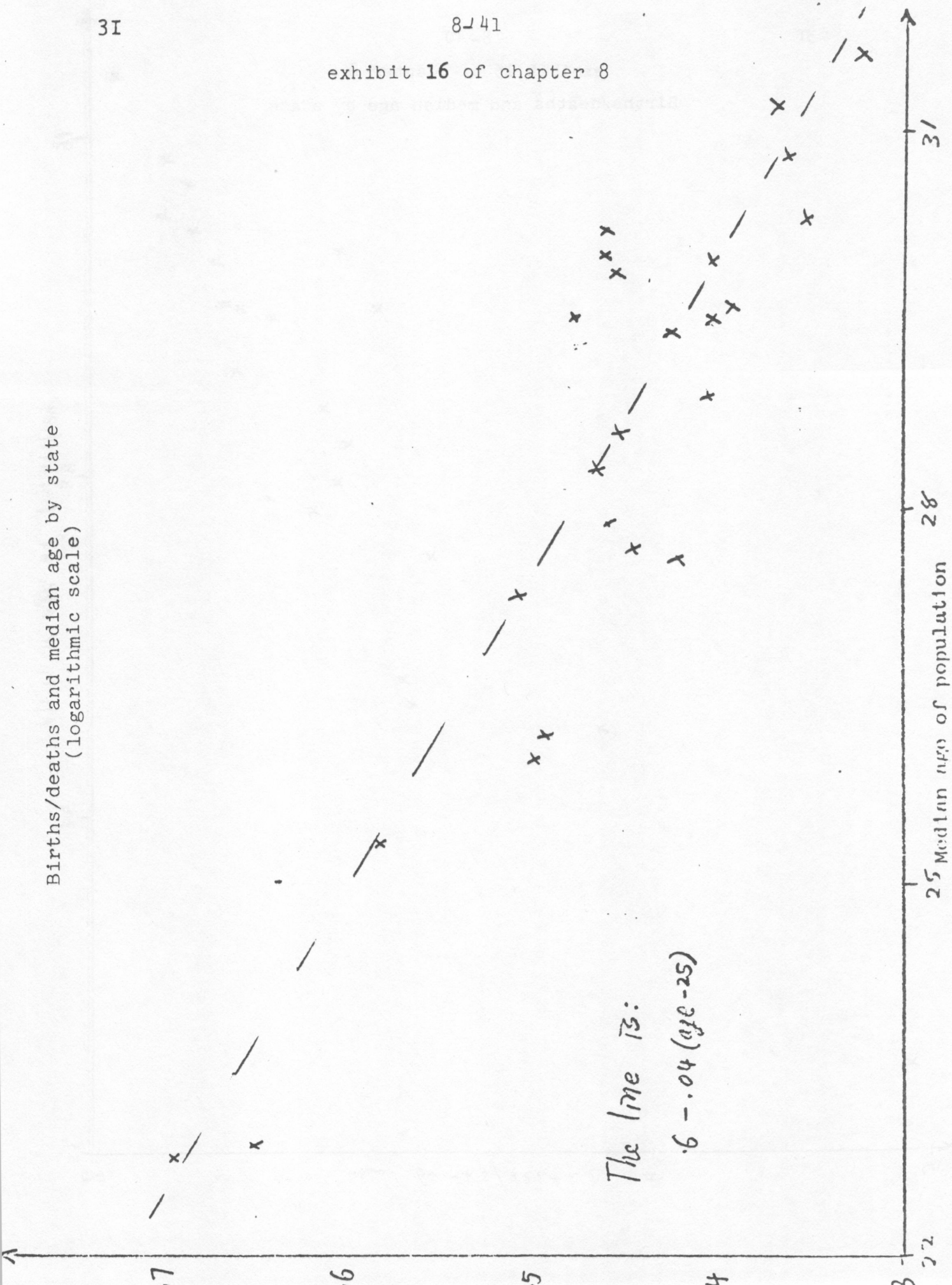
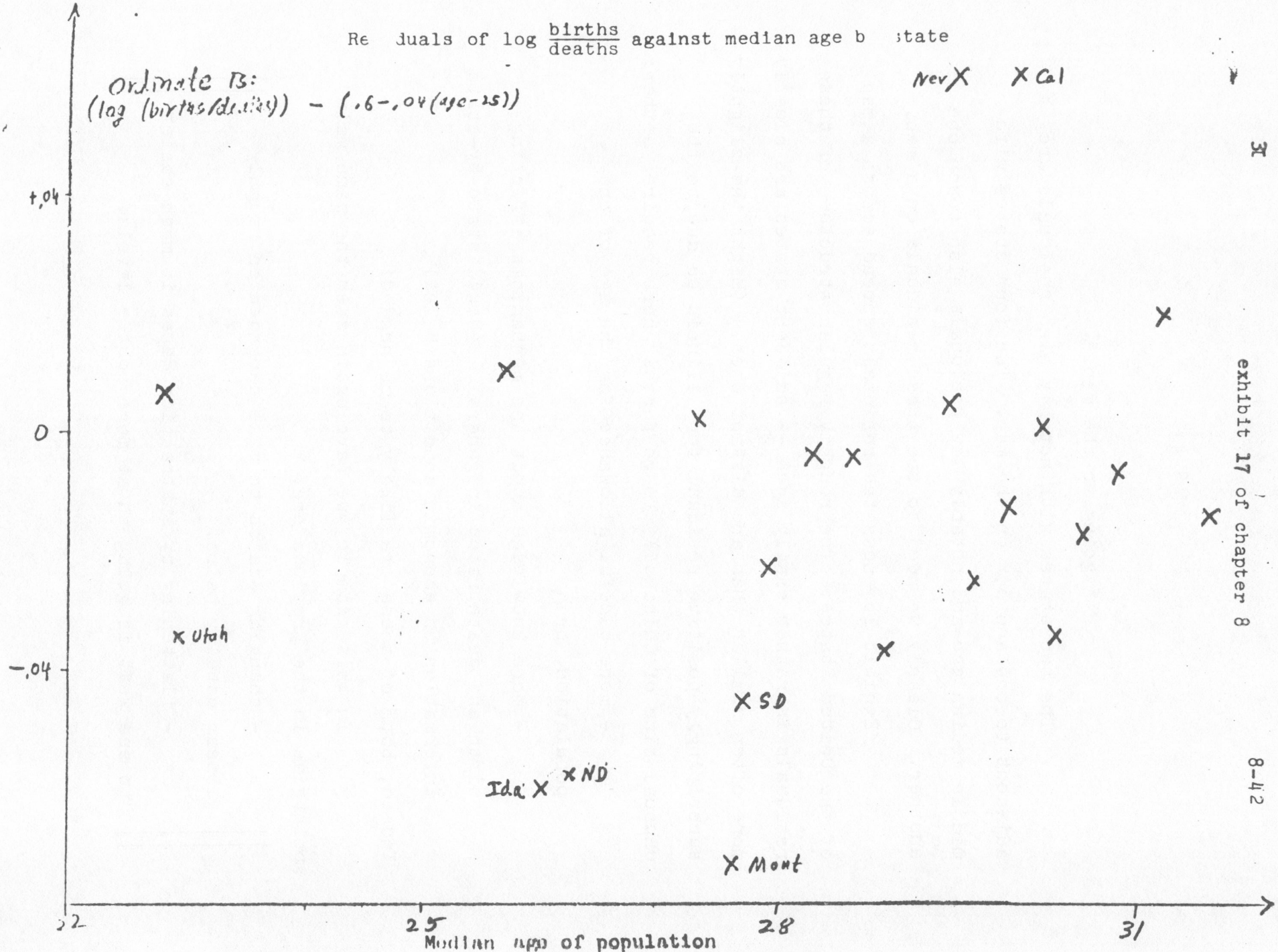


exhibit 16 of chapter 8

Births/deaths and median age by state
(logarithmic scale)



Re duals of log $\frac{\text{births}}{\text{deaths}}$ against median age b state



* going to the map *

The two states with notably high residuals are adjacent to one another on the map. The four states with middle median age and notably low residuals also touch one another. Clearly we need to see these residuals on a map.

Exhibit 18 shows the residuals spread across a map of the United States. The roughly regular structure of these residuals is rather easily seen -- adjacent states are clearly more often similar than are distant ones. (Might the slightly surprisingly positive residual for Illinois be due to the unusual size of Chicago?) To go further here requires either:

- more careful allowance for the age of the population, or
- more knowledge about the mechanisms affecting birth and death rates in general. (Would state-by-state information on economic conditions help?)

Indeed, both of these are likely to be needed!

In this example, we have again seen the same main points as in the previous one:

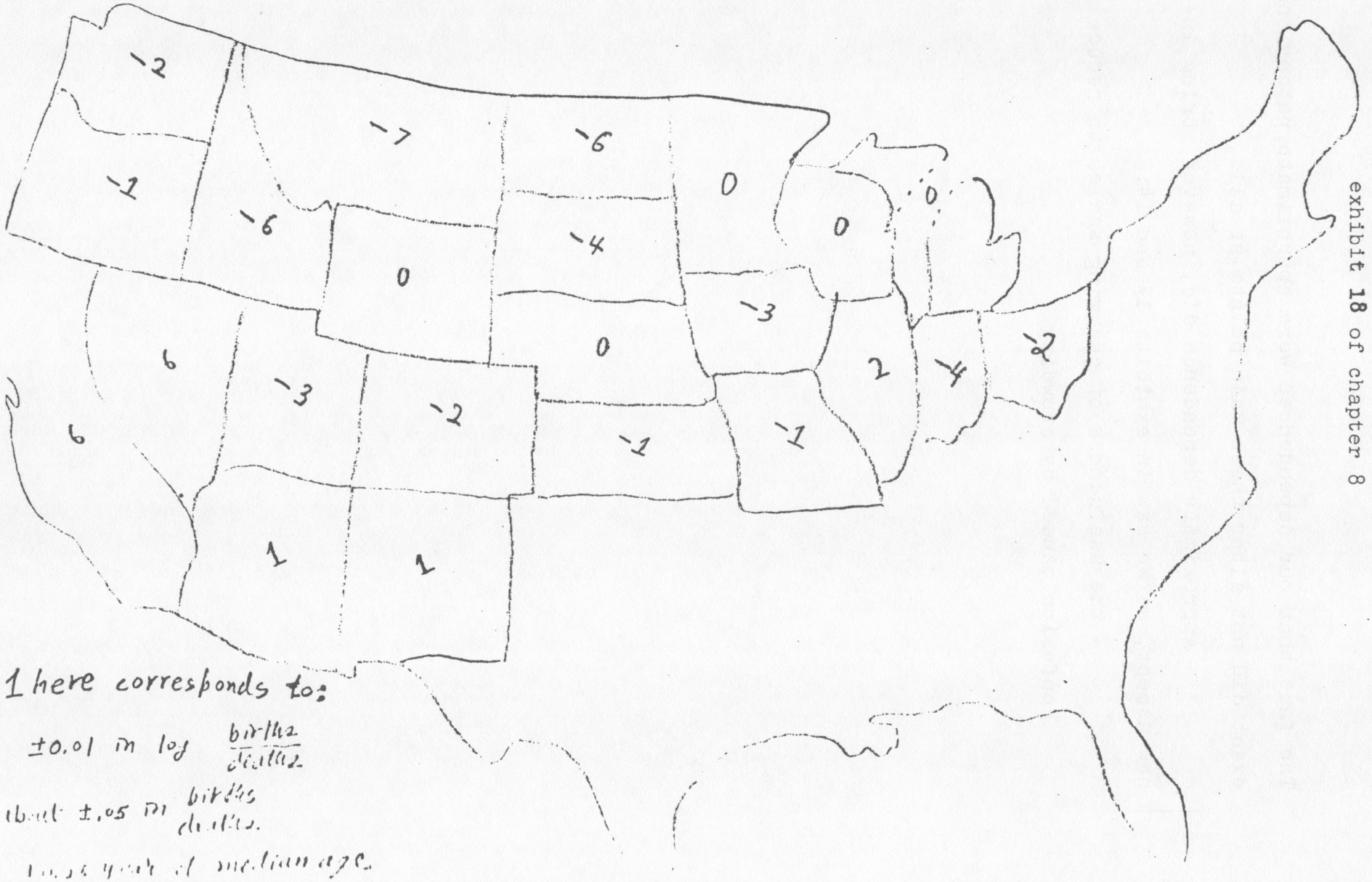
- changing scales to make dependences roughly linear usually helps.
- flattening by subtraction makes it much easier to see what is going on at more subtle levels.

The fact that our dependences were approximate rather than exact did not alter these main points at all.

Approximate dependence did, however, bring in one new aspect, as we saw in exhibits 13 and 14:

- the usefulness of changing scales to reduce confusion caused by crowding.

|| The residuals of exhibit 17 mapped ||



****

8H. How far have we come?

In this chapter we have met plots of y against x. Perhaps we have even come to know them a little.

The state of our progress is not measured by the specific techniques we have seen or understood -- though that kind of progress is essential. Rather, our progress is measured by our understanding, and calm acceptance, of such propositions as these:

1) graphs are friendly.

2) arithmetic often exists to make graphs possible.

3) graphs force us to note the unexpected; nothing could be more important.

4) different graphs show us quite different aspects of the same data.

5) there is no more reason to expect one graph to "tell all" than to expect one number to do the same.

6) comparison is not trivial; careful thought helps -- in graphical comparison as well as in arithmetic comparison.

7) "plotting y against x" involves significant choices -- how we express one or both variables can be crucial.

8) the first step in penetrating plotting is to straighten out the dependence or point spatter as much as reasonable.

9) plotting y^2, $\sqrt{y}$, log y, -1/y or the like in place of y is one plausible step to take in search of straightness.

10) plotting x^2, $\sqrt{x}$, log x, -1/x or the like instead of x is another.

11) once the plot is straightened, we can usually gain much by flattening it, usually by plotting residuals (with regard to the partial description implied by the straight line we may not have quite drawn in yet).

12) when plotting point spatters we may need to be careful about how we express x and y in order to avoid concealment by crowding.

In particular, we have learned -- or been reminded -- how to:

- understand about subtracting one curve from another.

- use two points to find an equation for a straight line.

*****

8P. Problems

1. Exhibits 19 to 22 contain selected information about the 15 smallest counties in North Carolina, South Carolina, Alabama and Georgia. (Source: 1962 City and County Data Book.) Plot y = % with less than five years schooling against x = % with less than $3000 income for at least two states. Continue the analysis. Comment.

2. The following data was obtained in preparing a standard curve for the determination of formaldehyde by the addition of chromatropic acid and concentrated sulphuric acid and the reading of the consequent purple color on a Beckman Model DU Spectrophotometer at 570 mμ.

Amount of CH_2O Used	Optical Density
0.1	0.086
0.3	0.269
0.5	0.446
0.6	0.538
0.7	0.626
0.9	0.782

Analyze graphically, using at least two graphs. Comment. (Bennett & Franklin, p. 216; from Roberts)

3. The relation between the amount of β-erythrodine dissolved in water and the turbidity of the solution -- as read on a colorimeter -- is not quite as simple. Some data gives:

exhibit 20 of chapter 8

South Carolina's 15 Smallest Counties

		Population (1960)				Families (1959)	Local Gov't (1957)	
(*)	Name	Total	% less than 5 yrs school**	% at least 65	per sq. mile	% less than $3000 income	% budget for education	Total budget ($1000's)
2424	McCormick	8,629	28.2	8.8	23	59.8	62.0	566
2130	Allendale	11,362	34.1	7.9	27	60.1	63.1	1082
2051	Jasper	12,237	37.1	7.5	19	60.2	79.7	1082
2050	Calhoun	12,256	26.1	7.9	33	68.2	76.6	659
1852	Saluda	14,554	17.0	9.7	33	50.6	68.9	656
1763	Edgefield	15,735	23.1	8.0	33	55.3	70.1	785
1720	Bamberg	16,274	26.6	8.3	41	58.5	63.1	1340
1622	Hampton	17,425	31.8	7.2	31	58.0	57.8	1312
1608	Barnwell	17,659	23.2	7.4	32	47.5	82.9	1565
1405	Fairfield	20,713	30.8	7.8	30	54.2	56.8	1596
1356	Abbeville	21,417	23.0	8.7	42	42.1	56.4	1717
1339	Lee	21,832	31.7	6.4	53	68.6	75.8	1671
1211	Dorchester	24,383	23.1	6.6	43	49.6	67.9	1684
1092	Colleton	27,816	29.3	7.5	27	57.9	46.7	3090
1066	Marlboro	28,529	29.6	6.8	59	58.3	63.4	1948
	(whole state)	((34,262))	(20.3)	(6.3)	(79)	(39.5)	( )	

(*) National rank by population.

** Of those 25 or over.

(( )) Median.

State has 46 counties.

exhibit 21 of chapter 8

Georgia's 15 Smallest Counties

(*)	Name	Population (1960) Total	% less than 5 yrs school**	% at least 65	per sq. mile	Families (1959) % less than $3000 income	Local Gov't (1957) % budget for education	Total budget ($1000's)
3077	Echols	1,876	28.0	8.5	4	55.6	59.5	122
3051	Quitman	2,432	34.8	9.3	14	70.0	79.5	205
3028	Glascoch	2,672	35.6	56.4	19	61.1	67.9	252
2979	Webster	3,247	34.5	8.7	17	71.2	73.3	225
2978	Schley	3,256	23.0	10.3	20	67.4	69.7	221
2970	Taliaferro	3,370	27.9	13.4	17	68.5	48.8	391
2950	Dawson	3,590	22.4	8.6	17	64.7	71.3	293
2934	Long	3,874	22.7	7.5	10	60.1	52.0	408
2875	Towns	4,538	11.3	10.2	27	63.7	33.4	724
2874	Baker	4,543	33.2	8.9	15	74.1	56.4	622
2873	Clay	4,551	26.8	10.5	20	66.6	59.2	417
2832	Lanier	5,097	30.1	7.5	31	57.4	64.5	686
2810	Charlton	5,313	26.5	6.2	7	44.0	49.1	703
2808	Heard	5,333	22.7	10.7	18	56.4	57.1	652
2805	Wheeler	5,342	29.4	9.5	18	63.7	54.4	418
	(whole state)	((12,038))	(17.6)	(7.4)	(68)	(35.6)	(45.0)	--

(*) National rank by population.

** Of those 25 or over.

(( )) Median.

State has 159 counties.

exhibit 22 of chapter 8

Alabama's 15 Smallest Counties

(*)	Name	Population (1960) Total	% less than 5 yrs school**	% at least 65	per sq. mile	Families (1959) % less than $3000 income	Local Gov't (1957) % budget for education	Total budget ($1000's)
2201	Coosa	10,726	18.4	10.7	17	51.5	39.5	1174
2179	Cleburne	10,911	21.5	9.4	19	52.4	48.2	1088
2034	Clay	12,400	14.4	13.0	21	54.0	48.3	1347
1945	Bullock	13,462	32.7	11.7	22	69.4	42.8	1727
1934	Greene	13,600	38.0	9.8	21	74.0	57.9	1514
1876	Lamar	14,271	14.7	11.0	24	51.4	37.5	2507
1870	Bibb	14,357	22.7	9.6	23	54.4	54.7	1405
1828	Winston	14,858	17.3	9.8	24	53.8	55.7	1072
1824	Crenshaw	14,909	23.9	11.2	24	69.5	51.4	1332
1796	Henry	15,286	27.8	9.0	27	63.8	56.0	1313
1794	Washington	15,372	23.3	7.8	14	51.7	54.5	1353
1790	Lowndes	15,417	37.2	9.2	22	72.1	55.4	1243
1729	Fayette	16,148	16.9	10.8	26	54.7	43.0	1425
1715	Cherokee	16,303	16.6	8.9	27	49.1	53.7	1498
1628	Perry	17,358	28.7	10.5	24	69.2	49.4	1689
	(whole state)	((25,738))	(16.3)	(8.0)	(64)	(39.1)	(45.8)	--

(*) National rank by population.

** Of those 25 or over.

(( )) Median.

State has 67 counties.

Concentration (in mg/ml)	Colorimeter reading
40	69
50	175
60	272
70	335
80	390
90	415

Analyze graphically. Comment. (Bennett & Franklin, p. 217, from Wislawski).

4. The amount of carbon in a clay can be measured directly by heating the clay until all the carbon compounds are burned, collecting the carbon dioxide thus formed, and measuring its amount. The amount of carbon can be estimated by combining the amounts of its constituents in a suitable standard way. The results of such measurements on 36 clays from South Devonshire, England are given in exhibit 23. Analyze graphically. Comment.

5. Exhibit 24 gives the % Democratic vote in each of 24 northeastern and central states for 12 presidential elections, 1920 to 1964. Plot the following:

(a) 1964 against 1956

(b) 1960 against 1920

(c) 1952 against 1932

(d) any one against any other that you think will have a close relationship.

exhibit 23 of chapter 8

Carbon Content of 36 Clays Measured Directly and Estimated Indirectly

Clay #	Direct Measurement	Indirect Estimate
1	1.53	2.46
2	0.87	1.54
3	0.28	0.70
4	0.27	-0.40
5	3.07	4.82
6	0.25	0.30
7	0.25	0.64
8	0.29	0.78
9	0.12	0.12
10	1.50	2.36
11	1.31	2.14
12	0.31	0.08
13	0.14	-0.01
14	2.98	4.53
15	6.84	9.94
16	2.15	3.68
17	1.35	1.84
18	0.40	0.97
19	4.18	6.14
20	0.22	0.52
21	0.38	0.40
22	0.24	0.46
23	1.79	2.80
24	0.58	2.09
25	6.55	9.68
26	2.54	4.08
27	1.43	2.80
28	2.74	3.93
29	6.08	8.22
30	0.75	0.28
31	0.16	0.35
32	5.06	7.49
33	0.86	1.41
34	0.16	-0.50
35	11.43	15.80
36	0.19	0.18

Source: C. A. Bennett and N. L. Franklin 1954. Statistical Analysis in Chemistry and the Chemical Industry. John Wiley, New York. Table 6.3 on page 213.

exhibit 24 of chapter 8

Percentage Democratic in 12 Presidential Elections for
24 Northeastern and Central States
(Percentage Democratic of Major Party Vote)

	1920	1924	1928	1932	1936	1940	1944	1948	1952	1956	1960	1964
Colorado	37.7	27.8	34.4	57.0	61.9	48.7	46.6	52.7	39.3	39.5	45.1	61.6
Connecticut	34.5	30.9	45.9	49.4	57.8	53.6	52.7	49.2	44.1	36.3	53.7	67.9
Delaware	43.0	38.9	33.9	48.8	54.9	54.8	54.6	49.4	48.1	44.7	50.8	61.1
Illinois	27.3	28.4	42.6	56.8	59.2	51.2	51.7	50.4	45.0	40.4	50.1	59.5
Indiana	42.3	41.2	39.9	56.0	57.5	49.3	47.1	49.6	41.4	39.9	44.8	56.2
Iowa	26.4	23.0	37.8	59.1	56.0	47.8	47.7	51.4	35.8	40.8	43.3	62.0
Kansas	33.4	27.7	27.3	54.8	53.9	42.7	39.4	45.4	30.7	34.3	39.3	54.6
Maine	30.2	23.3	31.1	43.6	42.8	48.8	47.5	42.7	33.8	29.1	43.0	68.8
Maryland	43.3	47.7	42.6	63.1	62.7	58.8	51.9	49.3	44.2	40.0	53.6	65.5
Massachusetts	28.9	28.5	50.5	52.1	55.1	53.4	52.9	55.9	45.6	40.5	60.4	76.5
Michigan	23.4	14.8	29.1	54.1	59.2	49.8	50.5	49.1	44.2	44.2	51.0	66.8
Minnesota	21.6	11.7	41.4	62.3	66.6	51.9	52.8	58.9	44.4	46.2	50.7	63.9
Nebraska	32.6	38.5	36.4	64.1	58.4	42.8	41.4	45.8	30.8	34.5	37.9	52.6
New Hampshire	39.7	36.7	41.2	49.3	50.9	53.2	52.1	47.1	39.1	33.9	46.6	63.9
New Jersey	29.6	30.6	40.0	51.0	60.1	51.8	50.7	47.7	42.5	34.6	50.4	66.0
New York	29.5	34.3	48.8	56.7	60.2	51.8	52.5	49.5	44.0	38.7	52.6	68.7
North Dakota	18.9	12.7	44.8	71.3	69.2	44.5	45.8	45.4	28.6	38.2	44.5	58.1
Ohio	39.8	28.9	34.7	51.5	60.8	52.2	49.8	50.1	43.2	38.9	46.7	62.9
Pennsylvania	29.2	22.6	34.2	47.1	58.2	53.5	51.4	48.0	47.0	43.4	51.2	65.2
Rhode Island	33.9	37.9	50.3	56.0	56.8	56.8	58.7	58.2	49.1	41.7	63.6	80.9
South Dakota	24.5	21.2	39.4	64.9	56.0	42.6	41.7	47.6	30.7	41.6	41.8	55.6
Vermont	23.5	16.7	33.0	41.6	43.4	45.1	42.9	37.5	28.3	27.8	41.4	66.3
West Virginia	43.9	47.1	41.3	55.1	60.7	57.1	54.9	57.6	51.9	45.9	52.7	67.9
Wisconsin	18.5	17.9	45.3	67.0	67.8	50.9	49.1	52.3	38.8	38.1	48.1	62.2

6. Determination of ethylene chlorohydrin. (264,270) translates as: "With 26.4 milligrams of ethylene chlorohydrin present, 27.0 milligrams were found". Data-sets (6): (264,270), (595,594), (1173,1183), (1777,1780), (2355,2370), (3578,3576). Source: K. Uhrig 1946. Determination of ethylene chlorohydrin. Industrial and Engineering Chemistry, Analytical Edition 18: 369 only. Table 1 on page 469. Problem: Choose a plot that is likely to be revealing by thinking hard. Explain the reasons for your choice. Make the plot.

7. Polarographic behavior of ions containing vanadium. (94,35) translates as: "For a concentration of vanadite ion of .094 millimoles per liter, the anodic diffusion constant in microamperes was 0.35 microamperes". Data-sets (8): (94,35), (278,98), (508,178), (880,309), (1548,563), (1840,696), (352,1285), (505,1813). (Last point: 5.05 millimoles/liter, 18.13 microamperes.) Source: J. J. Lingane 1945. Polarographic characteristics of vanadium in its various oxidation states. J. Amer. Chem. Soc. 67: 182-188. Table I on page 186. Problem: Make useful plots of diffusion current against vanadite concentration.

8. Amount of desired product in a chemical reaction after different reaction times and under different circumstances. (1;32,54;87,159,226) translates as: "In run 1, the amount of desired product in moles per liter was 0.032 after 80 minutes, 0.054 after 160 minutes, 0.087 after 320, 0.159 after 640, and

0.226 after 1280 minutes. Data-sets (16 runs under 16 different sets of conditions): (1;32,54;87,159,226), (2;147,234;343,342, 203), (3;48,108;225,346,420), (4;232,390;556,634,416), (5;37,38; 172,200,239), (6;179,283;405,342,216), (7;86,133;259,398,508), (8;309,514;722,764,389), (9;74,99;200,309,249), (10;253,353; 391,284,75), (11;133,271;430,580,494), (12;508,756;842,570,115), (13;96,158;276;339,230), (14;308,444;467,249,29), (15;228,372; 579,691,539), (16;626,880;895,434,58). Source: G. E. P. Box and W. G. Hunter 1962. A useful method for model-building. Technometrics 4: 301-318. Table 1 on page 304. Problem: Make useful plots of concentration at 640 minutes against concentration at 160 minutes.

9. Analysis of samples for chrysanthenic acid. (0,23) translates as: "When 0 micrograms of synthetic racemic chrysanthenic acid were added, the colorimeter scale reading was 23". Data-sets (13): (0,23), (5,32), (10,40), (20,54), (40,86), (60,118), (80,146), (100,179), (120,212), (140,240), (160,272), (180,300), (200,330). Source: A. A. Schreiber and D. B. McClellan 1954. Estimation of micro quantities of pyrethoids. Analytical Chemistry 26: 604-607. Table I on page 605. Problem: Make useful plots of colorimeter reading against amount of chrysanthenic acid.

10. Residual strength of 8-Oz cotton duck attacked by 4 different kinds of fungus. (3;97,105;103,101) translates as:

"After 3 hours of incubation, the strengths -- referenced to initial strength - 100 -- of the sample exposed to Thielaria was 97, that exposed to Humicola was 105, for Chaetomium was 103, for Myrothecium was 101." Data-sets (24): (3;97,105;103,101), (6;98,106;101,105), (9;95,107;99,95), (12;96,105;95,95), (15;97,106;90,100), (18;98,102;91,97), (21;97,101;78,98), (24;97,90;74,93), (27;90,81;71,82), (30;96,78;71,76), (33;89,73;65,67), (36;88,69;58,64), (39;89, 63;53,59), (42;86,59;47,54), (45;82,55;44,50), (48;79,53;44,42), (51;73,52;42,41), (54;73,41;40,40), (57;73,42;40,39), (60;68, 41;39,35), (63;59,36;38,37), (66;57,37;37,33), (69;57;31,35,34), (72;55,34;36,31). Source: E. Abrams 1950. Microbiological deterioration of cellulose during the first 72 hours of attack. Textile Research Journal 20: 71-86. Table 2 on page 75. Problem: Plot helpful curves for loss of strength from at least two kinds of fungus.

11. Make useful plots for one or more other pairs of reaction times. (Data-sets in problem 10 immediately above.)

12. Rapid analysis for caffeine. (257,131) translates as: "For a caffeine concentration of 0.257 milligrams in 100 milliliters, the average optical density was 0.131. Data-sets (20): (257,131), (498,262), (506,265), (514,263), (747,384), (760,393), (770,396), (996,512), (1013,518), (1027,523), (1245,633), (1266,643), (1284,650), (1494,760), (1519,768),

(1541,775), (1798,903), (2054,1040), (2311,1160), (2568,1290). Source: N. H. Ishler, T. P. Finucaine, and E. Borker 1948. Rapid spectrophotographic determination of caffeine. Analytical Chemistry 20: 1162-1166. Table 1 on page 1162. Problem: Make useful plots of optical density against caffeine concentration.

13. Survival in use by a public utility of automobiles and cars. (0h,990) translates as: "After 1/2 year, 0.990 of all vehicles were still in service." Data-sets (8): (0h,990), (1h,972), (2h,944), (3h,895), (4h,784), (5h,679), (6h,593), (7h,497). Source: S. A. Krane 1963. Analysis of survival data by regression techniques. Technometrics 5: 161-174. Table on page 168. His source, H. A. Cowles, Jr., 1957. Prediction of mortality characteristics of industriality property groups. Ph.D. Thesis, Iowa State University. Problem: Make helpful plots of fraction surviving against age.

14. Heat and entropy contents of a sodium silicate. (400,3080,885) translates as: "The increases from 'room temperature' (298.16°K) to an absolute temperature of 400°K were 3,080 calories per mole for the heat content of Wa_2SiO_3 and 8.85 calories/degree/mole for its entropy content." Data-sets (17): (400,3080,885), (500,6300,1604), (600,9650,221- , (700,13190,2760), (800,16910,3256), (900,20730,3708), (1000, 24700,4124), (1100,28770,4511), (1200,32940,4874), (1300,37210, 5216), (1361,39870,5416), (1361,52340,6332), (1400,54010,6453),

(1500,58290,6748), (1600,62570,7024), (1700,66850,7284), (1800,71130,7528). (The last point, at 1800°K gives a heat content change of 71,130 calories/mole and an entropy content change of 75.28 calories/degree/mole.) Source: B. F. Naylor 1945. High-temperature heat contents of sodium metasilicate and sodium disilicate. J. Amer. Chem. Soc. 67: 466-467. Table II on page 467. Problem: Make helpful plots of the increase in heat content against temperature.

15. For Naylor's data (immediately above) make helpful plots of increase in entropy content against temperature.

16. Equilibrium splitting of plutonium tribromide by water (gases at high temperature). (911,153) translates as: "For an absolute temperature of 911°K, the observed equilibrium constant was 0.0153/atmosphere." Data-sets (11): (911,153), (914,156), (919,149), (920,163), (882,246), (876,282), (875,247), (883,243), (815,704), (817,502), (816,692). Source: I. Shift and N. R. Davidson 19 . Equilibrium in the vapor-phase hydrolysis of plutonium tribromide. Paper 6.24, at pages 831-8 of The Transuranium Elements, edited by Seaberg, Katz, and Manning. National Nuclear Energy Series IV-14B. McGraw Hill. Table 2 on page 835. Problem: Make useful plots of equilibrium constant against temperature. Which three of the 11 data-sets do you think the authors rejected?

17. Sales of Swiss bond issues since World War II. (46,527) translates as: "In 1946, total sales of Swiss bonds -- governmental and private -- were 527 million francs." Data-sets (23): (46,527), (47,276), (48,472), (49,342), (50,174), (51,434), (52,333), (53,249), (54,242), (55,492), (56,613), (57,1148), (58,827), (59,686), (60,890), (61,1023), (62,1124), (63,2091), (64,2503), (65,2523), (66,2292), (67,2446), (68,2648). Source: Swiss Statistical Abstract, issued by the Swiss Credit Bank, November 1969. (Title also in French and German.) Table on page 46. Problem: Make useful plots based on the data from 1950 to 1968.

18. Comparison of two ways of measuring the water content of samples of the sea bed. (0 to 3;76,76) translates as: "For a sample from 0 to 3 inches below the surface of the sea level, measurement of % water by drying in an oven gave 76%, measurement by analyzing for chloride -- and using the known concentration of chloride in deep sea water -- gave 76%." Data-sets (14): (0 to 3;76,76), (3 to 6;68,72), (6 to 9;69,69), (9 to 12;67,67), (12 to 15;60,64), (15 to 18;62,62), (18 to 21;60,60), (21 to 24;58,59), (24 to 27,57,57), (27 to 30;55,56), (30 to 33;55,55), (33 to 36;55,55), (36 to 39;53,54), (39 to 42;54,54). Source: L. J. Anderson 1948. Conductometric titration of chloride in sea water and marine sediments. Analytical Chemistry 20: 618-619. Table II on page 619.

Problem: Find differences between water by chloride and water by oven. Make a stem-and-leaf display of these differences. Comment on the appearance of this display. Find and plot residuals from a straight-line fit of water content against depth, separately for each method of finding water content. What do you conclude about the two methods of measuring water?

19. Find two different collections of (x,y) points that interest you, and make useful plots.